HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Cohesin-dockerin interaction based method to facilitate fast domain shuffling of cellobiohydrolases

Eero A. Kiviniemi

Masters' thesis

University of Helsinki

Faculty of Agriculture and Forestry

Department of Microbiology

Biotechnology

04/2018

Työn nimi / Arbetets titel – Title

Cohesin-dockerin interaction based method to facilitate fast domain shuffling of cellobiohydrolases

Oppiaine /Läroämne – Subject

Biotechnology

Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
Masters' thesis	09.04.2018	83

Tiivistelmä/Referat – Abstract

Microbial cellulases, e.g. cellobiohydrolases, are able to degrade cellulose and lignocellulosic biomass to smaller glucose-containing monomers and oligomers. Cellulases are often multi-domain enzymes comprised of different protein domains (i.e. modules), which have different functions. The main two components, which often appear in cellulases, are the cellulose-binding module (CBM) and the catalytic domain. The CBMs bind to cellulose, bringing the catalytic domains close to their substrate and increasing the amount of enzymes on the substrate surface. The catalytic domain performs the cleavage of the substrate, e.g. in the case of cellobiohydrolases hydrolyses or "cuts" the crystalline cellulose chain into smaller soluble saccharides, mainly cellobiose.

Unlike aerobic fungi, which utilize free extracellular enzymes to break down cellulose, anaerobic microbes often use a different kind of strategy. Their cellulases are organized and bound to the cell surface in a macromolecular protein complex, the cellulosome. The core of the cellulosome is formed of a scaffolding protein (the scaffoldin) consisting mainly of multiple consecutive cohesin domains, into which the catalytic subunits of enzymes attach via a dockerin domain. This creates a protein complex with multiple different catalytic domains and activities arranged in close proximity to each other. Dockerins and cohesins are known to bind each other with one of the strongest receptor-ligand -pair forces known to nature. Dockerin containing fusion proteins have also been successfully combined *in vitro* with proteins containing their natural counterparts, cohesins, to create functional multiprotein complexes.

In this Master's thesis work the goal was to 1) produce fusion proteins in which different CBMs were connected to dockerin domains, 2) combine these fusions with cohesin-catalytic domain fusion proteins to create stable CBM and catalytic domain containing enzyme complexes, 3) to characterize these enzyme complexes in respect of their thermostability and cellulose hydrolysis capacity and 4) to ultimately create a robust and fast domain shuffling method for multi-domain cellobiohydrolases (CBH) to facilitate their faster screening. The hypothesis of the experiments was that different CBMs fused with a dockerin domain and the cellobiohydrolase catalytic domain fused with a cohesin domain could be produced separately and then be combined to produce a functional two-domain enzyme with a dockerin-cohesin "linker" in between. In this way time and work could be saved because not every different CBM- catalytic domain -pair would have to be cloned and produced separately.

Several CBM-dockerin fusion proteins (in which the CBM were of fungal or bacterial origin) were tested for expression in heterologous hosts, either in *Saccharomyces cerevisiae* or *Escherichia coli*. The purified proteins were combined with a fungal glycoside hydrolase family 7 (GH7) cellobiohydrolase-cohesin fusion protein produced in *S. cerevisiae*. The characterization of the catalytic domain-CBM -complexes formed through cohesin-dockerin interaction included thermostability measurements using circular dichroism and activity assays using soluble and insoluble cellulosic substrate. The results were compared to enzyme controls comprising of the same CBM and catalytic domain connected by a simple peptide linker. The results showed that the cohesin-dockerin – linked cellobiohydrolase complex performed in the cellulose hydrolysis studies in a similar manner as the directly linked enzyme controls at temperature of 50˚C and 60 ˚C. At temperatures of 70 ˚C the complex did not perform as well as the control enzymes, apparently due to the instability of the dockerin-cohesin interaction. The thermostability measurements of the enzymes, together with the previously published data supported the hydrolysis results and this hypothesis. The future work should be aimed at enhancing the thermostability of the cohesin-dockerin interaction as well as on verifying the results on different cellulase fusion complexes.

Avainsanat – Nyckelord – Keywords

Cellulase, cellobiohydrolase, enzyme, cellulose-binding domain, cellulose hydrolysis, dockerin, cohesin, fusion protein, domain swapping/shuffling, screening, protein engineering, biotechnology

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

The thesis work was conducted at VTT, Espoo, Finland, under the supervision of Sanni Voutilainen, Anu Koivula and Kiyohiko Igarashi from VTT and Marko Virta from the University of Helsinki.

Työn nimi / Arbetets titel – Title

Cohesiini-dockeriini -vuorovaikutuspohjainen menetelmä sellobiohydrolaasien nopean domeenien vaihtamisen mahdollistamiseksi

Oppiaine /Läroämne – Subject

Biotekniikka

Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
Maisterin tutkielma	09.04.2018	83

Tiivistelmä/Referat – Abstract

Mikrobien sellulaasit kuten sellobiohydrolaasit pystyvät hajottamaan selluloosaa ja lignoselluloosaista biomassaa pienemmiksi, glukoosia sisältäviksi monomeereiksi ja oligomeereiksi. Sellulaasit ovat usein monidomeenientsyymejä, jotka koostuvat erilaisista proteiinidomeeneista (tai -moduuleista), joilla on eri toimintoja. Sellulaasien pääkomponentteina ovat usein selluloosaan sitoutuva domeeni (cellulose binding module, CBM) ja katalyyttinen domeeni. CBM:t sitoutuvat selluloosaan tuoden katalyyttiset domeenit lähelle substraattiaan ja näin nostaen entsyymien määrää substraatin pinnalla. Katalyyttiset domeenit pilkkovat substraatin ja esimerkiksi sellobiohydrolaasien kohdalla hydrolysoivat tai katkaisevat kiteisen selluloosaketjun pienemmiksi, liukoisiksi sakkarideiksi, pääasiassa sellobioosiksi.

Toisin kuin aerobiset sienet, jotka käyttävät vapaita solunulkoisia entsyymejä hajottaakseen selluloosaa, anaerobiset mikrobit käyttävät usein toisenlaista strategiaa. Niiden sellulaasit ovat järjestäytyneet ja sitoutuneet solun pinnalle makromolekulaariseksi kompleksiksi nimeltään sellulosomi. Sellulosomin ydin muodostuu scaffoldin nimisestä proteiinista, joka koostuu pääasiassa useista perättäisistä cohesiini domeeneista, joihin katalyyttiset alayksiköt kiinnittyvät dockeriini domeeniensa välityksellä. Näin muodostuu proteiinikompleksi, jossa useita erilaisia katalyyttisiä domeeneja ja aktiivisuuksia järjestäytyy lähelle toisiaan. Dockeriinien ja cohesiinien tiedetään sitoutuvan toisiinsa yhdellä vahvimmista reseprotien ja ligandien välisistä voimista, joita luonnosta tunnetaan. Dockeriinin sisältäviä fuusioproteiineja on myös onnistuneesti yhdistetty niiden luonnollisen vastinparin, cohesiinin, sisältävien proteiinien kanssa *in vitro* tuottaen toimivia moniproteiinikomplekseja.

Tässä maisterin tutkielman työssä tavoitteena oli 1) tuottaa fuusioproteiineja, joissa erilaisia CBM:iä oli yhdistetty dockeriini domeeneihin, 2) yhdistää nämä fuusioproteiinit cohesiini-katalyyttinen domeeni fuusioproteiinien kanssa luoden vakaita CBM:n ja katalyyttisen domeenin sisältäviä entsyymikomplekseja, 3) karakterisoida näitä entsyymikomplekseja niiden lämmönkestävyyden ja selluloosan hydrolyysikyvyn suhteen ja 4) lopulta luoda vakaa ja nopea domeenienvaihtometodi sellobiohydrolaaseille näiden nopeamman skriinaamisen mahdollistamiseksi. Kokeen hypoteesina oli se, että eri CBM:iä fuusioituna dockeriinien kanssa ja sellobiohydrolaasien katalyyttisiä domeeneja fuusioituna cohesiinien kanssa voitaisiin tuottaa erikseen ja yhdistää jälkikäteen tuottaen toimivia ”kaksidomeenisia” entsyymejä cohesiini-dockeriini ”linkkerillä”. Menetelmän ansiosta aikaa ja työtä voitaisiin säästää entsyymien skriinaamisessa kun kaikkia erilaisia CBM-katalyyttinen domeeni -pareja ei tarvitsisi kloonata ja tuottaa erikseen.

Useiden CBM-dockeriini fuusioproteiinien (joissa CBM:t olivat joko sieni- tai bakteeriperäisiä) heterologista ekspressiota testattiin eri isäntäorganismeissa, joko *Saccharomyces cerevisiae* –hiivassa tai *Escherichia coli* –bakteerissa. Puhdistetut proteiinit yhdistettiin hiivassa tuotetun sieniperäisen glykosidihydrolaasi perhe 7:n sellobiohydrolaasi-cohesiini fuusioproteiinin kanssa. Cohesiini-dockeriini vuorovaikutuksen voimasta muodostuneen katalyyttinen domeeni-CBM -kompleksin karakterisointiin kuului lämpövakausmittauksia sekä aktiivisuusmäärityksiä käyttäen liukoista ja liukenematonta substraattia. Tuloksia verrattiin kontrollientsyymeihin, jotka koostuivat samasta CBM:stä ja katalyyttisestä domeenista, joita yhdisti yksinkertainen linkkeripeptidi. Tulokset osoittivat, että cohesiini-dockeriini vuorovaikutuksen avulla muodostettu sellobiohydrolaasikompleksi toimi selluloosan hydrolyysikokeissa yhtäläisesti peptidilinkkerillä varustetun entsyymikontrollin kanssa 50:n ja 60 ˚C:n lämpötiloissa. Kuitenkaan 70 ˚C:n lämpötilassa kompleksi ei toiminut enää yhtä hyvin kuin kontrollientsyymi, ilmeisesti cohesiini-dockeriini vuorovaikutuksen epävakaudesta johtuen. Entsyymien lämpövakausmittaukset aiemmin julkaistujen tuloksien kanssa tukivat hydrolyysituloksia ja tätä hypoteesia. Tuleva tutkimus tulisi tähdätä parantamaan cohesiini-dockeriini vuorovaikutuksen lämpövakautta ja vahvistamaan tulokset eri sellulaasifuusiokomplekseilla.

Avainsanat – Nyckelord – Keywords

Sellulaasi, sellobiohydrolaasi, entsyymi, selluloosaan sitoutuva domeeni, selluloosan hydrolyysi, dockeriini, cohesiini, fuusioproteiini, domeenien sekoittaminen / vaihtaminen, skriinaus, proteiinimuokkaus, biotekniikka

Muita tietoja – Övriga uppgifter – Additional information

Tämä maisterin tutkielman työ suoritettiin VTT:llä, Espoossa, VTT:n Sanni Voutilaisen, Anu Koivulan ja Kiyohiko Igarashin sekä Helsingin yliopiston Marko Virran ohjaamina.

Table of Contents

Abbreviations

CAZy = The Carbohydrate-Active enZYmes database

CBH = Cellobiohydrolase

CBM = Cellulose-binding module

CD = Circular dichroism

CipA = Cellulosomal scaffoldin subunit, scaffoldin

Ct = *Clostridium thermocellum*

CV = Column volume

DocD = *Clostridium thermocellum* endoglucanase D dockerin

DocS = *Clostridium thermocellum* Cel48S (CelS) dockerin

EG = Endoglucanase

GH (6/7) = Glycoside hydrolase (family 6/7)

IPTG = Isopropyl β-D-1-thiogalactopyranoside

O/N = Overnight

PAHBAH = Para-hydroxybenzoic acid hydrazine

RT = Room temperature

Sc = *Saccharomyces cerevisiae*

Te = *Talaromyces emersonii*

TMW = Theoretical molecular weight

Tr = *Trichoderma reesei*

v/v = Volume per volume

w/v = Weight per volume

1. Introduction

Cellulases are used in diverse industrial applications for example in textile industry, food industry, in detergent manufacturing, pulp and paper industry and in the processes of making biofuels from lignocellulosic feedstock (Zhang et al. 2006). Especially the production of ethanol from lignocellulosic feedstock has recently gathered much interest which has also accelerated the research on cellulases (Payne et al. 2015). Plant polysaccharides, cellulose and hemicellulose, account for more than 50% of all plant biomass and are therefore the most abundant organic molecules found on land (Gilbert and Hazlewood 1993), which also contributes to their enormous potential in using them as a source of renewable energy or a feedstock for the production of ethanol and other added value products and chemicals. Fungi are considered to be efficient degraders of lignocellulosic biomass, and they have evolved to produce a number of different types of lignocellulolytic enzyme batteries with varying enzymatic compositions to deal with this recalcitrant substrate. Because of their significant activities and readiness to be produced in large scales (over 100 g/L in industrial expression hosts), the fungal cellulases have been considered to be an excellent biotechnological tool for utilizing the potential of lignocellulosic biomass (Payne et al. 2015).

In order to be able to apply the enzymes in industrial applications, such as bioethanol production, the price of the enzymes needs to be affordable. As such, however, the enzymatic hydrolysis is still a major bottleneck and a cost factor in the production (Viikari et al. 2012). Improving the enzymes' thermostabilities, pH tolerance, specific activities, binding on the substrate and reducing of the end-product inhibition would make the production of bioethanol from cellulosic feedstock more cost efficient (Zhang et al. 2006, Viikari et al. 2012, Voutilainen et al. 2014). "Domain shuffling" of enzymes, i.e. combining different domains (i.e. modules) of enzymes such as cellobiohydrolases or xylanases from different origins to create single fusion proteins with enhanced properties, has also been show to work efficiently and it has been proposed to be a useful tactic to "reach optimum enzymes for application purposes" (Voutilainen et al. 2014).

Lignocellulosic biomass consist mainly of cellulose, hemicellulose and lignin, that are packed tightly around each other forming a hard to access bundles of recalcitrant material resistant to enzymatic attack. Cellulose is the major component forming crystalline, insoluble fibrous structures within it. Many different kind of enzymes, in addition to chemical or physical pretreatments are needed in

order to degrade lignocellulosic biomass to its single sugar constituents (Viikari et al. 2012). In addition, compositions of the cellulosic feedstocks and the pretreatment methods affect to the composition of the optimal enzyme cocktail needed to be applied for efficient biomass degradation in each situation. Thus it would be desirable to be able to quickly construct various enzymes and enzymes cocktails with different activities and screen these cocktails on the selected pretreated biomass. In addition, it would be also beneficial to be able to design and produce, not just one specific enzyme, but rather a battery of different cellulolytic enzymes with a multitude of affinities and activities in order to deal with the heterogeneous nature of the lignocellulosic feedstock and the demanding conditions of enzymatic hydrolysis.

In order to study whether the screening of multi-domain cellulases could be carried out in a faster way, a method to facilitate fast domain shuffling of multi-domain cellobiohydrolases was tested in this thesis. In the method the catalytic domains and the CBMs are produced separately from each other as fusion protein with bacterial cellulosomal protein domains, dockerin and cohesin, to facilitate their combining afterwards to functional multi-domain enzymes. The method has previously been shown to work in domain shuffling of a natively dockerin bearing bacterial endoglucanase catalytic domain with different cohesin and cellulose-binding modules (CBMs) bearing fusion proteins (Carrard et al. 2000). In this thesis the goal was particularly to test whether the method could be applied to cellobiohydrolases, by testing whether several different dockerin-CBM fusions could be produced and the catalytic domain and CBM containing cellobiohydrolase complexes be constructed and characterised.

1.1. Free fungal cellulases

Glycoside hydrolases (GH) are enzymes that hydrolyse the glycosidic bonds of various carbohydrates (CAZy, http://www.cazy.org/, AFMB laboratory, 2017). Cellulases are a non-uniform group of these glycoside hydrolases that specifically hydrolyse the β-1,4-O-glycosidic bonds between two glucose residues of the cellulose chain (Gilbert and Hazlewood 1993). Fungal cellulases are often multidomain enzymes which comprise of a catalytic domain connected to a cellulose-binding module by a flexible linker region (Figure 1) (Payne et al. 2015). The most studied fungal cellulolytic system is that of *Trichoderma reesei*, an aerobic, filamentous fungus that is also used as an industrial expression host for various (hemi)cellulolytic enzymes (Viikari et al. 2012).

Protein domains have their own folds and independent functions in relation to the rest of the protein (Artzi et al. 2017) and they can be classified into different protein families. Cellulase domains are classified in the Carbohydrate-Active enZYmes database (CAZy: http://www.cazy.org/) into different families based on their sequence similarity and 3D fold. There are currently more than 450 000 different catalytic domains classified into over 140 GH families. In addition there are more than 100 000 carbohydrate-binding domains classified into over 80 families in the CAZy database. The 3D structures of the linker peptides have not been able to be solved, and are thus not classified into any protein families. The linker sequences connecting the different fungal cellulase domains are usually rich in serine and threonine amino acids, are partially O-glycosylated, and may vary vastly in length (Gilbert and Hazlewood 1993, Viikari et al. 2012).

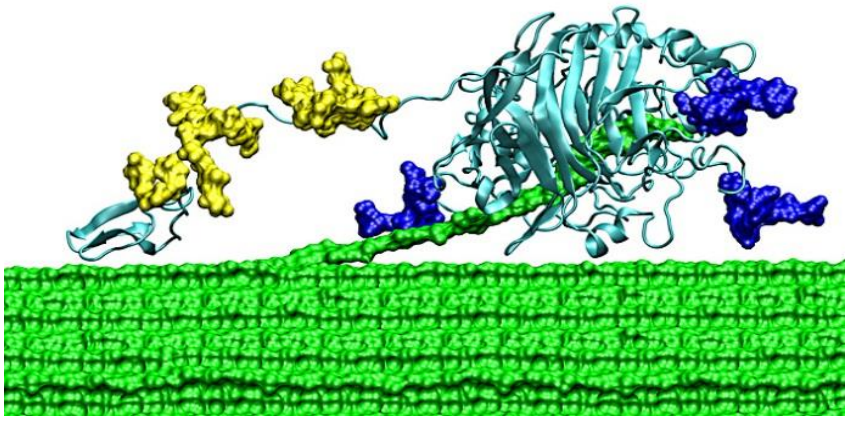


Figure 1: *Trichoderma reesei* Cel7A cellobiohydrolase (CBH I). CBH I is the major cellulolytic enzyme produced by the filamentous model fungus *T. reesei*. It is a 2-domain enzyme, composed of a catalytic domain (belonging to GH7 family) and a family-1 cellulose–binding domain or module (CBM1). The two domains are connected by an approximately 40 amino acids long linker peptide, which is partially O-glycosylated by *T. reesei*. The active site of this cellobiohydrolase is located in a tunnel made of surface loops and containing altogether 10 binding subsites for the glucose units of a cellulose chain to attach to. Figure by G. Beckham (NREL, 2018).

At least ten different enzymes are needed for the complete hydrolysis of lignocellulosic carbohydrates, of which more than six are considered essential, depending a little on the substrate (Viikari et al. 2012). These enzymes include cellobiohydrolases (CBHs) (EC 3.2.1.91), endoglucanases (EC 3.2.1.4), different xylanases (EC 3.2.1.8) and β-glucosidases (EC 3.2.1.21), as well as lytic polysaccharide monoxygenases (LPMOs) and acetyl xylan esterases (EC 3.1.1.6) (Viikari et al. 2012). The different cellulolytic enzymes also act synergistically with each other, e.g. the endoglucanases cut the cellulose chains from the middle, on the more amorphous region of the cellulose, creating free chain-ends for the cellobiohydrolases to attach to. The recently discovered LPMO enzymes are

oxidative enzymes able to cleave crystalline cellulose, thus also creating new chain-ends to CBHs, even on the crystalline part of the cellulose. The CBHs can then processively hydrolyse the crystalline cellulose chain to cellobiose units. The CBHs are particularly important enzymes in crystalline cellulose degradation and the most abundant enzyme in the commercially available enzyme mixtures. The β-glucosidases cleave the soluble cellobiose to two glucose units, thus preventing end-product inhibition of cellobiohydrolases (Gilbert and Hazlewood 1993). Xylanases and other accessory (or auxiliary) activity enzymes depolymerize xylan (and other lignocellulose polymers), making the cellulose surfaces also more accessible for the cellulases.

The fungal CBHs belong to two different GH families, GH6 (CBH II) and GH7 (CBH I). The major CBH responsible for the cellulose degradation of the model fungus *T. reesei* is the GH7 family CBH (CBH I/ Tr Cel7A). The 3D structures of several fungal GH7 family cellobiohydrolases have been determined (www.cazy.org). The structures have revealed a unique active site tunnel structure where protein surface loops create a tunnel for the substrate to enter, containing several binding subsites for the glucose units of the cellulose chain to interact with, along with the actual hydrolytic active site. The extensive hydrogen bonding networks and stacking of the aromatic residues at the binding subsites with the sugar rings facilitate the strict substrate recognition, binding and subsequent hydrolysis of the substrate by the catalytic residues (usually glutamic and/or aspartic acids) to mainly cellobiose (Divne et al. 1994, Ståhlberg et al. 1996, Grassick et al. 2004). The hydrolysis is performed at the reducing chain end of the substrate by the retaining hydrolysis mechanism (Knowles et al. 1988, Divne et al. 1994, Ståhlberg et al. 1996). The long active site tunnel of the CBH I with the sugar binding subsites facilitates the retention of the substrate close to the active site even after the cleavage of the cellobiose enabling the consecutive processive movement of the CBH I along the same cellulose chain, without disassociation in the middle. For comparison the CBH II active site tunnel is much shorter than the CBH I active site tunnel, which also makes the action of the enzyme less processive (Divne et al. 1994). Also the hydrolysis mechanism of the enzyme is inverting instead of retaining (Knowles et al. 1988). Furthermore, although the GH7 family endoglucanases (EGs) have similar overall folds in their catalytic domains as the CBH I, they lack the long loops forming the active site tunnels and thus do not exhibit such processive mode of action (Viikari et al. 2012, Payne et al. 2015). These different modes of action allow the enzymes to degrade the cellulosic substrate in different ways and at different locations and to complement each other for an enhanced synergistic degradation of cellulosic biomass (Divne et al. 1994, Viikari et al. 2012).

The carbohydrate-binding domains have different affinities to different substrates and can bind e.g. cellulose, chitin, starch, xylans or mannans. The carbohydrate binding domains are currently classified into some 80 different families, of which domains reported to be able to recognize and bind cellulose i.e. cellulose-binding modules (CBM) are found at least in 21 families (http://www.cazy.org). The different CBMs also can bind to either crystalline or amorphous cellulose and have different affinities towards them (Carrard et al. 2000, Voutilainen et al. 2014) and are accordingly further divided into 3 functional types (A, B and C) based on their substrate binding specificities (Boraston et al. 2004). The differential binding enables the enzymes with different CBMs to degrade cellulose more efficiently in cooperation with each other (Carrard et al. 2000).

The importance of the CBMs for the cellulose hydrolysis and even their role and function in substrate degradation has been a long standing source of debate in the scientific community. Many roles for the carbohydrate-binding domains have been described (Guillén et al. 2010), and even an active role or involvement of the CBMs in the disruption of the cellulosic substrate has been suggested on several occasions (Caspi et al. 2008, Guillén et al. 2010). The main role for most CBMs, nevertheless, seems to be to attach the enzymes to the substrate surface and to keep the catalytic domains in close proximity to their substrate (Guillén et al. 2010, Várnai et al. 2014, Voutilainen et al. 2014). The importance of the CBMs for the cellulose hydrolysis in industrially relevant conditions (i.e. under high biomass consistency) has also been questioned recently (Várnai et al. 2014). Nevertheless, the CBMs have been traditionally considered as essential parts of the cellulases to achieve efficient cellulose hydrolysis rates and to enable the effective plant biomass degradation in nature. In addition some of the CBMs have been observed to enhance the thermostability and the overall stability of the enzymes especially during long periods of hydrolysis, in the presence of denaturing agents and/or at high temperatures (Várnai et al. 2014, Voutilainen et al. 2014).

Although their substrate specificities vary widely, the carbohydrate-binding domains most often seem to be composed of β-sheet structures forming different folds (and carbohydrate-binding families) (Guillén et al. 2010). The first 3D structure of a CBM was determined already in 1989 for the CBM (CBM1) of Tr Cel7A (CBH I) by Kraulis et al. (Figures 1 and 2 a). The CBM1 is the smallest carbohydrate-binding domain composed of only 36 amino acids and forming a wedge-shaped 3D structure composed of β –strands. Of the different CBM tertiary structures the β-sandwich –fold is the most common, occurring for example in the carbohydrate binding domain families 2 and 3 (Figure 2 b and c)(Boraston et al. 2004, Guillén et al. 2010). The β-sandwich consists of two β-sheets

that have from three to six antiparallel β-strands in them. Most or nearly all CBMs with a β-sandwich fold contain at least one structural metal ion in their structure (Figure 2 b), that is not in most cases directly involved in the ligand binding (Boraston et al. 2004).

The carbohydrate-binding domains attach to their substrate mainly via the conserved aromatic residues located at their ligand recognition surfaces, utilizing the strong van der Waals interactions forming between them and the sugar rings of the substrate. In addition some of the polar amino acid side-chains located at the recognition surface may too interact with the sugars by hydrogen bonding, reinforcing the interaction even further. With functional type A carbohydrate-binding domains, such as CBMs able to bind crystalline cellulose, the aromatic residues form a "flat hydrophobic surface", which is able to interact with the flat surface of the water insoluble crystalline cellulose (Figure 2)(Boraston et al. 2004, Guillén et al. 2010). With carbohydrate-binding domains, which bind amorphous cellulose or xylan (type B), the ligand recognition surface is not as planar, but forms more of a cleft where the interaction between the substrate recognition residues and free polysaccharide chains happens (Boraston et al. 2004). The aromatic side chains' orientation "forms twisted or sandwich platforms" and slight changes in the orientation of the residues or the topology of the cleft determines the substrate or specific oligosaccharides that the residues are able to interact with i.e. the substrate (recognition) specificity of the domain. The same reason also explains why different carbohydrate binding domains with similar structures recognize different substrates. The third type (type C) CBM domains are consequently only able to bind mono-, di-, or trisaccharides because they have steric restrictions in their substrate binding sites which prevent the recognition of larger oligo-, or polysaccharides (Boraston et al. 2004, Guillén et al. 2010).

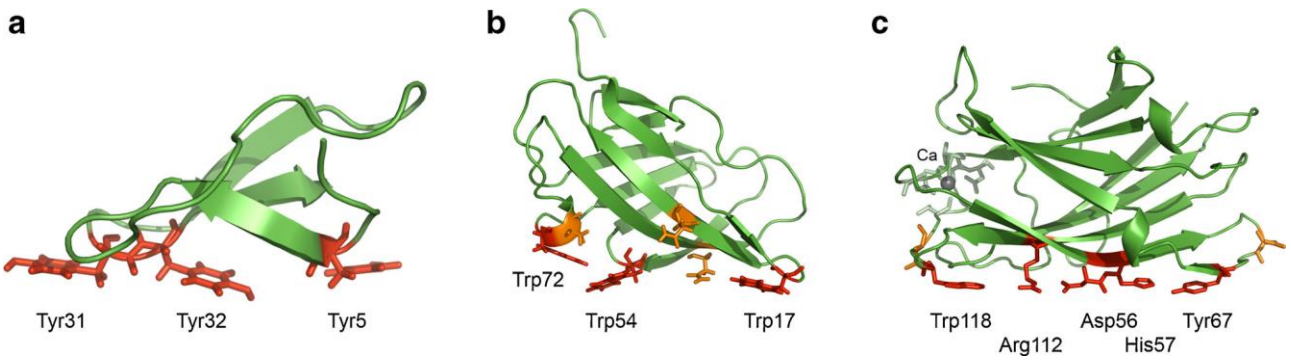


Figure 2. 3D structures of three different (type A) CBMs, belonging to the carbohydrate-binding families 1 (a), 2 (b) and 3 (c). a) CBM of *Trichoderma reesei* CBHI (Tr Cel7A)(PDB ID 2CBH). b) CBM of *Cellulomonas fimi* Xyn10A. c) CBM3 of *C. thermocellum* CipA (PDB ID 1NBC_A). The CBMs in Figures 2 a and 2 c correspond to the CBMs used in this thesis. Figure taken from Voutilainen et al. (2014).

In nature the carbohydrate-binding domains can be found attached to the catalytic domains in different arrangements and also in different ratios (Guillén et al. 2010). The specificities of the CBMs and the catalytic domains they are attached to often reflect each other (Guillén et al. 2010). The best combinations of CBMs and the catalytic domains from a biotechnological point of view may not, however, always be readily available in the same enzyme. Engineering of these enzymes has often been found useful to enhance the enzymes' properties for different applications (Carrard et al. 2000, Zhang et al. 2006, Viikari et al. 2012, Voutilainen et al. 2014). For example addition of a bacterial cellulosomal scaffoldin subunit (CipA) CBM (CBM3) from a thermophilic bacteria *Clostridium thermocellum* (Ct) to a thermostable single-domain GH7 family cellobiohydrolase (Cel7A) of *Talaromyces emersonii* (Te) has been shown to enhance both the thermostability of the enzyme and its cellulose hydrolysis capability at high temperatures (Voutilainen et al. 2014). For comparison, the newly engineered enzyme (TeCel7A-CBM3) was several times more efficient at hydrolyzing cellulose at high temperatures than the industrially important and much used cellobiohydrolase I (CBH I) enzyme of *Trichoderma reesei* (Tr) (Figure 1)(Voutilainen et al. 2014).

Thermostable enzymes are especially interesting from an industrial point of view since enhancing the thermostability of the enzymes confers to the enzymes' longer lifetimes and higher specific activities at high temperatures. This results to lowered needs of enzymes in hydrolysis, but also allows enhanced flexibility to process configurations due to higher solubility of the substrate, lowered viscosity, increased diffusion rates and the possibility to use different processes and process configurations e.g. in simultaneous saccharification and fermentation or consolidated bioprocesses in bioethanol production (Viikari et al. 2012, Artzi 2017). Thus screening for enzyme thermostability can be considered as one of the critical steps in enzyme engineering and performance enhancement of the enzymatic hydrolysis of cellulose.

1.2. Bacterial cellulosome complex and the cohesin-dockerin interaction

1.2.1. The cellulosome

Cellulosomes are the cellulose breaking machineries of many anaerobic bacteria (and possibly some anaerobic fungi) found in different demanding environments such as soil, hot springs or human and animal guts, where they live in colonies or communities of mixed species of other cellulolytic and non-cellulolytic microbes (Bayer et al. 1994, Fanutti et al. 1995, Stahl et al. 2012, Artzi et al. 2017). The cellulosome of the thermophilic anaerobic bacteria *Clostridium thermocellum* (Figure 3) has

been studied the most and has served as an archetype for cellulosomes and their research (Smith and Bayer 2013).

The main principle functions for the cellulosome are the adherence of the cellulosome and the bacteria to the substrate surface, to serve as a platform for the enzymatic subunit domains to attach to and to finally degrade the cellulosic substrate with the attached cellulolytic enzymes. The cellulosomes consist of a large variety of different proteins and enzymes, arranged in a close proximity with each other to enable enhanced synergistic action on the lignocellulosic substrate (Bayer et al. 1994). The proteins and enzymes are transported to the outer cell surface where they then self-assemble to a macromolecular complex the size of ~100 nm and more than 2 MDa that is attached to the cell surface (Lamed et al. 1983, Stahl et al 2012). The expression of the different cellulosomal components or different catalytic subunits of *C. thermocellum* is determined by the composition of the substrate and its sugar components (Artzi et al. 2017), which enables the bacteria to fine tune the compositions of the cellulosomes according to the environmental needs. As the cells grow, mature and proliferate, the cellulosomes may also be released into the extracellular matrix to continue the degradation independently (Bayer et al. 1994).

The core of the cellulosome consist of one large scaffolding subunit, which has been termed “the scaffoldin” (CipA, for cellulose integrating protein)(Figure 3)(Lamed et al. 1983, Gerngross et al. 1993, Bayer et al. 1994). The scaffoldin in itself consist mainly of many consecutive repeating domains of conserved sequence and structure called the cohesins (Shimon et al. 1997), which are connected to each other via flexible linker regions (Smith and Bayer 2013). The role of the cohesin domains in the scaffoldin is to serve as binding sites for different catalytic subunits, such as cellulolytic enzymes, that arrange along the cellulosome in a random order. The enzymatic subunits attach to the cohesins tightly via their special docking domains or so called dockerins with an affinity that is among one of the highest ligand-receptor pair forces known to nature (Fanutti et al. 1995, Pagès et al. 1997, Stahl et al. 2012). The architecture or compositions of the cellulosomes and the numbers of cohesin domains in the scaffoldins vary highly between different species of bacteria (Pagès et al. 1997, Artzi et al. 2017). For example the scaffoldin of C*. thermocellum* cellulosome is approximately the size of 210 kDa (Lamed et al. 1983) and contains nine conserved cohesin domains in it (Gerngross et al. 1993). The numbers of different dockerin bearing enzymatic subunits in a single organism can rise even up to hundreds (Artzi et al. 2017). Since the catalytic subunits attach to the scaffoldin in a random order, the compositions of different cellulosomes within one species or even

a single individual can differ highly from one another, yielding cellulosomes with varying distinct catalytic activities. By this diverse assembly system, the different catalytic domains are brought in a close proximity to each other and the microbial surface to produce synergistic effects in respect of complex lignocellulosic substrate hydrolysis and to minimize the hydrolysis produced sugars' diffusion away from the microbe (Artzi et al. 2017).

Another critically important part of the cellulosome is the cellulose-binding domain of the scaffoldin. These CBMs, usually belonging to the family 3 of CBMs, like the CBM of *C. thermocellum* CipA (CBM3, Figures 2 c and 3), attach the cellulosome to the surface of the cellulose, which brings the catalytic subunits close to their substrate and enables the bacteria to attach to different cellulosic surfaces (Artzi et al. 2017). As a type A CBM the CBM3 has affinity especially for crystalline cellulose (Boraston et al. 2004).

The cellulosomal systems of different organisms vary vastly between species and the different components they contain in them are numerous. In addition to cohesins and CBMs the aforementioned scaffoldins or "primary" scaffoldins can also contain a varying number of other domains such as hydrophilic domains and another types of dockerin domains (as with *Clostridium thermocellum*: a type-II dockerin), that attach the scaffoldins to the cell surface or other scaffoldins (Pagès et al. 1997, Smith and Bayer 2013, Artzi et al. 2017). The type-II dockerin (of *C. thermocellum*) differs from the type-I dockerins of the cellulases structurally and sequentially and does not bind to the cohesins in the primary scaffoldin subunit, but attaches itself to a type-II cohesin domain of the cell surface anchoring proteins or anchoring scaffoldins via a type-II cohesin-dockerin interaction (Lytle et al. 1996). The primary scaffoldin of *C. thermocellum* also contains an 'X-module' close to the type-II dockerin domain. The role of the X-module is not yet thoroughly understood but it has been suggested to confer mechanical stability to the dockerin domain and thus to enhance the affinity of the type-II cohesin-dockerin interaction (Schoeler et al. 2014, Artzi et al. 2017). The anchoring proteins or anchoring scaffoldins (Figure 3 B) can contain either a single cohesin or multiple cohesin domains, which respectively can facilitate either one or more of these primary (or other) scaffoldins (Artzi et al. 2017). The architectures of the different cellulosomal structures can be really elaborate. The cellulosomes can for example contain different sorts of anchoring proteins and "adaptor" scaffoldins that facilitate the binding of yet another degree of primary scaffoldins to the cell surface. This enables the formation of multicellulosomal enzyme complexes that have been estimated to bear even as much as 160 enzymatic subunits in them. Cellulosomes can also be found

in free extracellular solution conformations in which the scaffoldins are not bound to the cell surface. The multiplicity and versatile biological use of these elaborate systems confers to the usability of the systems from a biotechnological point of view which is why the use of designer cellulosomal systems has been considered to be a very promising and also already a usable technology in different sorts of applications such as in protein engineering (Bayer et al. 1994, Artzi et al. 2017, Bayer 2017).

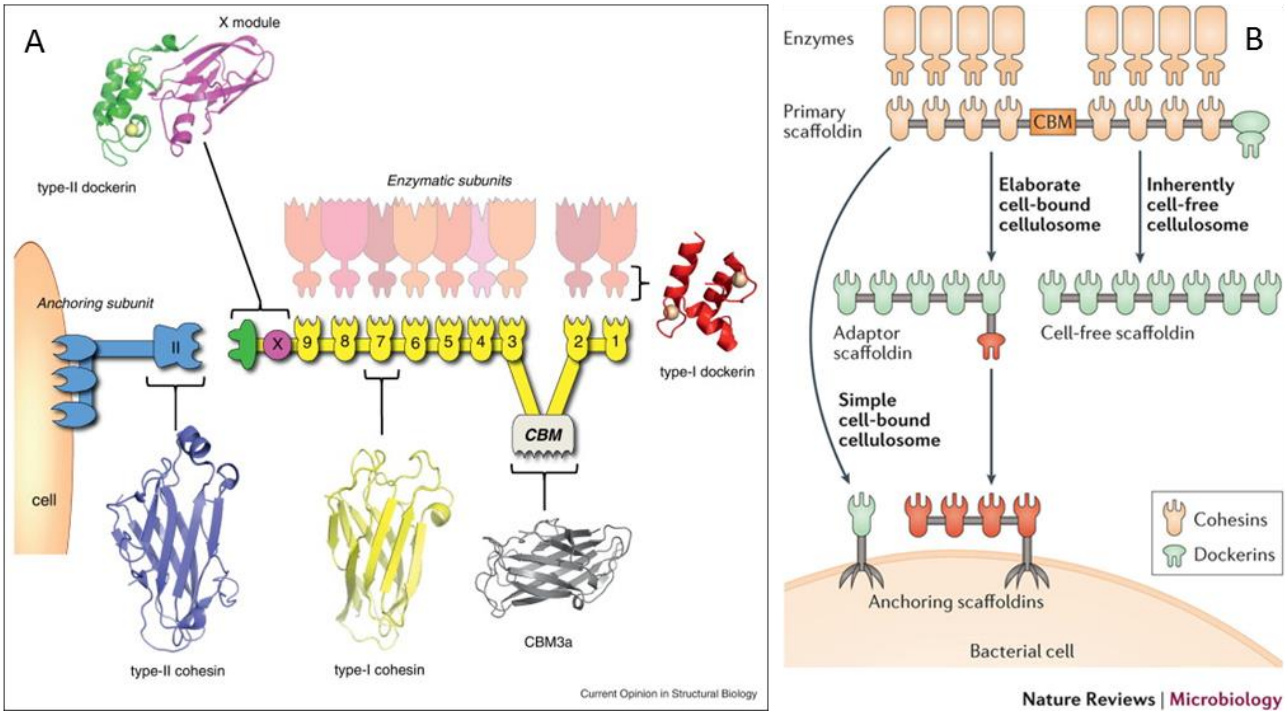


Figure 3. Different types of scaffoldins. A) The basic cellulosomal system of *Clostridium thermocellum*. The cellulosome is built around the non-catalytic cellulosomal subunit, the scaffoldin. The *C. thermocellum* scaffoldin consist of nine type-I cohesin subunits (yellow), a family 3a cellulose-binding domain (CBM, grey), an X-module (purple) and type- II dockerin (green) that attaches the scaffoldin to the type-II cohesin domain of the cell surface bound anchoring protein or subunit (blue) via a type-II interaction. The enzymatic subunits attach to the cellulosome in a random order via their (C-terminal) type-I dockerin domains (red) that exhibit high affinity for the type-I cohesins of the scaffoldin subunit (Smith and Bayer 2013). B) Depicted is the cellulosomal system and different types of cellulosomal and free scaffoldins of *Clostridium clariflavum*. Enzymes attach directly to the cohesins of the primary scaffoldin via a type-I cohesin-dockerin interaction. The primary scaffoldin can attach either directly to the cell surface anchoring scaffoldin or to an adaptor of cell free scaffoldin via a type-II cohesin dockerin interaction. The adaptor scaffoldins and also some of the anchoring scaffoldins can hold several primary scaffoldins or adaptor scaffoldins in them respectively amplifying the number of catalytic enzyme domains attached to the cell surface in close proximity to each other by several degree compared to primary scaffoldins that attach directly to the single cohesin bearing anchoring scaffoldins. Matching cohesin-dockerin pairs are depicted by matching colors (Artzi et al. 2017).

1.2.2. Cohesin-dockerin interaction

Bacterial cellulosomal proteins, dockerin and cohesin are known to bind each other with one of the strongest receptor-ligand pair forces known to nature, with mechanical rupture forces of >120 pN needed to break the bond and a dissociation constant of $<10^{-11}$ M (Mechaly et al. 2001, Stahl et al. 2012). The rupture forces measured are similar to those of the streptavidin-biotin pair, which is known for its high affinity and applications arising therefrom (Wilchek and Bayer 1999), when the forces are loaded at the same rate for each of the pairs (Florin 1994, Merkel 1999, Stahl et al. 2012).

Mechanical rupture forces even as high as 600 to 750 pN have been reported for the X-module-dockerin-cohesin interaction of *Ruminococcus flavefaciens*, with evidence of the X-module contributing to the stability of the interaction (Schoeler et al. 2014). The rupture forces were approximately half of the rupture force of a covalent bond between a gold and a thiol group making the interaction "one of the strongest bimolecular interactions reported", although the rupture forces did not however correlate with the biochemical affinity of the interaction ($K_D >10^{-8}$ M).

The high affinity of the (type-I) cohesin-dockerin interaction is based on the very conservative structure of both the cohesin and the dockerin (Pagès et al. 1997, Lytle et al. 2000). In the interface between the two domains multiple conserved amino acids act with each other, creating the strong bond, comprising of many, both hydrophobic and hydrogen bonding, interactions (Figures 4 and 5)(Carvalho et al. 2003, Mechaly et al. 2000, Mechaly et al. 2001, Karpol et al. 2008, Stahl et al. 2012, Jobst et al. 2015).

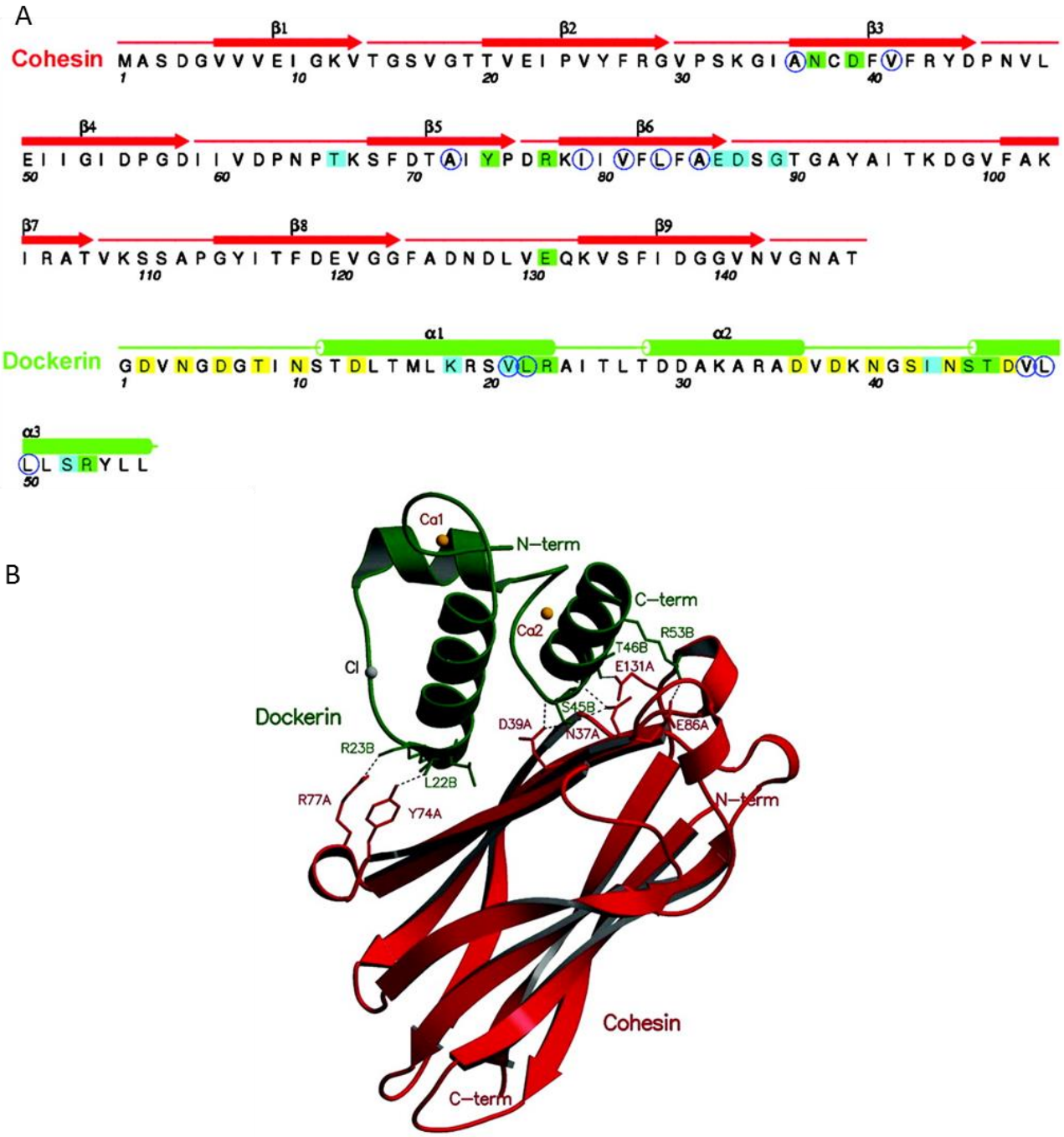


Figure 4. The interactions between Ct CipA cohesin2 and the dockerin domain of Ct xylanase 10B and the residues responsible for them. A) The amino acid sequences of the domains and interacting residues. Highlighted are the residues responsible for the binding of the dockerin to the cohesin domain in just one of its binding modes. The green squares show the residues that are in direct contact between the two domains, the blue squares depict the interactions mediated by bridging water molecules and the blue circles show the hydrophobic residues located at the cohesin-dockerin contact surface. Shown in the figure are also the secondary structures of the cohesin and dockerin (red arrows and green cylinders) and the calcium binding residues of the dockerin (yellow squares). B) A 3D-model of the interaction between the two domains. On green is shown the model of the dockerin domain and on red the model of the cohesin domain. Depicted are the main residues responsible for the (direct) interaction between the two domains in the same binding mode as in Figure A as well as the two structural calciums of the dockerin domain and the termini of the two domains. Figures taken from Carvalho et al. 2003.

Nearly all the type-I dockerin domains found in the cellulosomal cellulases bind to the CipA cohesin domains of the same species predominantly non-selectively, although some exceptions to this rule and variations in the affinities between different pairs of the domains do exist (Yaron et al. 1995, Lytle et al. 1996, Carvalho et al. 2003). The type-II dockerin domain of the CipA, however does not have affinity for the type-I cohesins of the CipA and does not bind to them (Lytle et al. 1996). The affinities of the interaction of cohesins and dockerins between different bacterial species vary highly and for example the affinities of the domains in the thermophilic organism *C. thermocellum* are much higher (by over 10-fold) than those of a mesophilic organism *Clostridium cellulolyticum* (Mechaly et al. 2001).

The type-I cohesin domains are approximately 140 amino acids long proteins (Figure 4)(Gerngross et al. 1993, Shimon et al. 1997). They have a secondary structures that consist mainly of β-sheets and their interconnecting loops that together form "nine-stranded β-sandwiches with a jelly-roll topology" type of structures (Figures 3 and 4)(Shimon et al. 1997). Of the nine cohesin domains of the Ct CipA (from here on CipA only refers to the *C. thermocellum* scaffoldin) most show homology and sequence identities of 96% to 100% with each other with two or three showing sequence identities which are lower than those (69% identity at lowest) (Gerngross et al. 1993, Lytle et al. 1996). The structures of the cohesins are recalcitrant as the second cohesin domain of the CipA has been reported to be stable at even as high temperatures as 85˚C (Lytle et al. 1996). In addition the cohesin domains have been said to "possess a compact structure that is highly resistant to proteases and to denaturing agents" (Miras et al. 2002).

In the cohesin domain the interacting area and the residues involved in the binding of dockerin are located only on one specific location on one side of the β-barrel, where residues mainly from the β-strands 3, 5 and 6 interact with the dockerin domain (Figure 4)(Miras et al. 2002, Carvalho et al. 2003). Also the alanine-94 residue in Figure 4 A of the cohesin has been shown to be part of the recognition between the two domains by Miras et al. (2002), in addition to the rest of the residues shown in the figure. The interacting face of the cohesin domain is mostly negatively charged (Carvalho et al. 2003). No apparent ligand binding cleft can be found in the structure of the cohesin domain but the interaction is mediated by exposed surface residues (Figure 4 A and B) (Shimon et al. 1997, Carvalho et al. 2003).

The dockerin domain is an approximately 70 amino acid long protein domain that contains two ~25 amino acid long duplicated sequential F-hand structures of a calcium binding loop followed by an α-

helix. The loop-helix motifs of the dockerin resemble but still clearly differ from the typical eukaryotic calcium binding helix-loop-helix structure (i.e. an EF-hand motif) by lacking the first helix of the structure (Figure 4)(Pagès et al. 1997, Lytle et al. 2000, Jobst et al. 2015). The duplicated sequence confers high internal symmetry to the dockerin domain and in some of the dockerins the two duplicated loop-helix motifs even possess nearly perfect symmetry to one another (Carvalho et al. 2003, Jobst et al. 2015). This symmetry contributes to the double binding mode activity of dockerins to cohesins, allowing the domain to interact with the cohesin domain in two possible orientations in which the dockerin is rotated ~180˚ when compared to the other one (Figure 5)(Carvalho et al. 2003, Carvalho et al. 2007, Jobst et al. 2015). No apparent favoring of either of the binding modes can be seen for dockerins with the consensus sequence binding residues (Figure 5 A) (Jobst et al. 2015). However, when residues of one binding mode are mutated, the other binding mode prevails (Carvalho et al. 2007, Jobst et al. 2015). This characteristic of the binding modes could be used when designing protein systems where a more precise orientation of the proteins is needed (Jobst et al. 2015).

The structure of the calcium-binding loops as of the rest of the dockerin is highly conserved (Lytle et al. 2000, Carvalho et al. 2003, Jobst et al. 2015). The binding of the calcium to the structure is coordinated by the side chains of the five conserved asparagine/ aspartic acids residues either directly or through hydrogen bonding mediated by a water molecule, and by the backbone oxygen of a less conserved residue (Ser/Thr/Arg/Lys). The same residues coordinate the calcium binding at the same positions in both of the duplicated sequences (Figure 5)(Carvalho et al. 2003, Jobst et al. 2015). The calcium is needed for the correct domain folding (Lytle et al 2000) and for the formation of the high affinity interaction between cohesin and dockerin (Pagès et al. 1997, Craig et al. 2006, Stahl et al. 2012). Withdrawal of the calcium by the addition of EDTA results in the dissociation of the cohesin and dockerin from each other (Craig et al. 2006), but the binding can be re-established by the removal of EDTA and addition of Ca^{2+} (Stahl et al. 2012). This reversible, but high affinity binding of the two counterparts and the dissociation in mild non-denaturing conditions by the addition of EDTA can be utilized in different applications such as immunodetection and affinity purification of natural cellulosomal and recombinant proteins (Craig et al. 2006).

Previously it had been considered that the dockerin undergoes a conformational change upon binding its cohesin counterpart, while the cohesin remained unchanged (Carvalho et al. 2003). Only recently it was discovered, that this does not hold true for the dockerin domain and the domain

remains unchanged when it binds to cohesin (Chen et al. 2014). It has been, however, witnessed that changes in the structures or rearrangement of the residues and the interaction surfaces of at least some cohesin and dockerin pairs (such as the *R. flavefaciens* CttA dockerin and cell surface anchoring protein cohesin) do happen when the complex is subjected to mechanical force or pulling, during which the interaction tightens by a catch bond mechanism (Schoeler et al. 2014).

The interaction between the cohesin and dockerin domain is highly specific and many conserved binding residues in both of the domains are of critical importance for the binding affinity. Even a single amino acid mutation at these positions has been show to significantly reduce the binding affinity (Mechaly et al. 2001, Handelsman et al. 2004, Karpol et al. 2008, Stahl et al. 2012). In addition, although the dockerin-cohesin recognition is highly species specific, single amino acid mutations in the dockerin domain have been shown to change the specificity, allowing the dockerins to recognize their counterparts from another species in addition to the ones of the same species (Mechaly et al. 2001, Handelsman et al. 2004). In the dockerin domain especially the residues located in or near the loops at the ends of the helixes 1 and 3 are important for the binding of the dockerin to the cohesin (Figures 4 and 5)(Carvalho et al. 2003, Jobst et al. 2015. In addition to the "actual" binding residues presented in Figures 4 and 5, also the calcium binding residues of the dockerin are essential for the interaction (Karpol et al. 2008). In order for the dockerin to remain its high affinity for the cohesin at least one of its calcium binding loops (and the calcium binding residues within it) needs to remain intact (Karpol et al. 2008). In addition, the two consecutive binding or "recognition" residues within the calcium binding loop (ST or SS in *C. thermocellum*, Figure 5) were found to be among the most crucial residues for the high affinity (and species specificity) of the interaction (Mechaly et al. 2000, Mechaly et al. 2001). Mutations even in only one of these residues (the latter threonine /serine) resulted in a reduced intraspecies "self-recognition" by more than three orders of magnitude (Mechaly et al. 2001). Furthermore, mutation of four of the other recognition residues in each of the duplicated sequences of the dockerin (altogether eight residues mutated) completely eliminated the affinity between the domains (Jobst et al. 2015). Similar to dockerin, corresponding "critical" binding residues have also been identified in the cohesin domain (Handelsman et al. 2004).

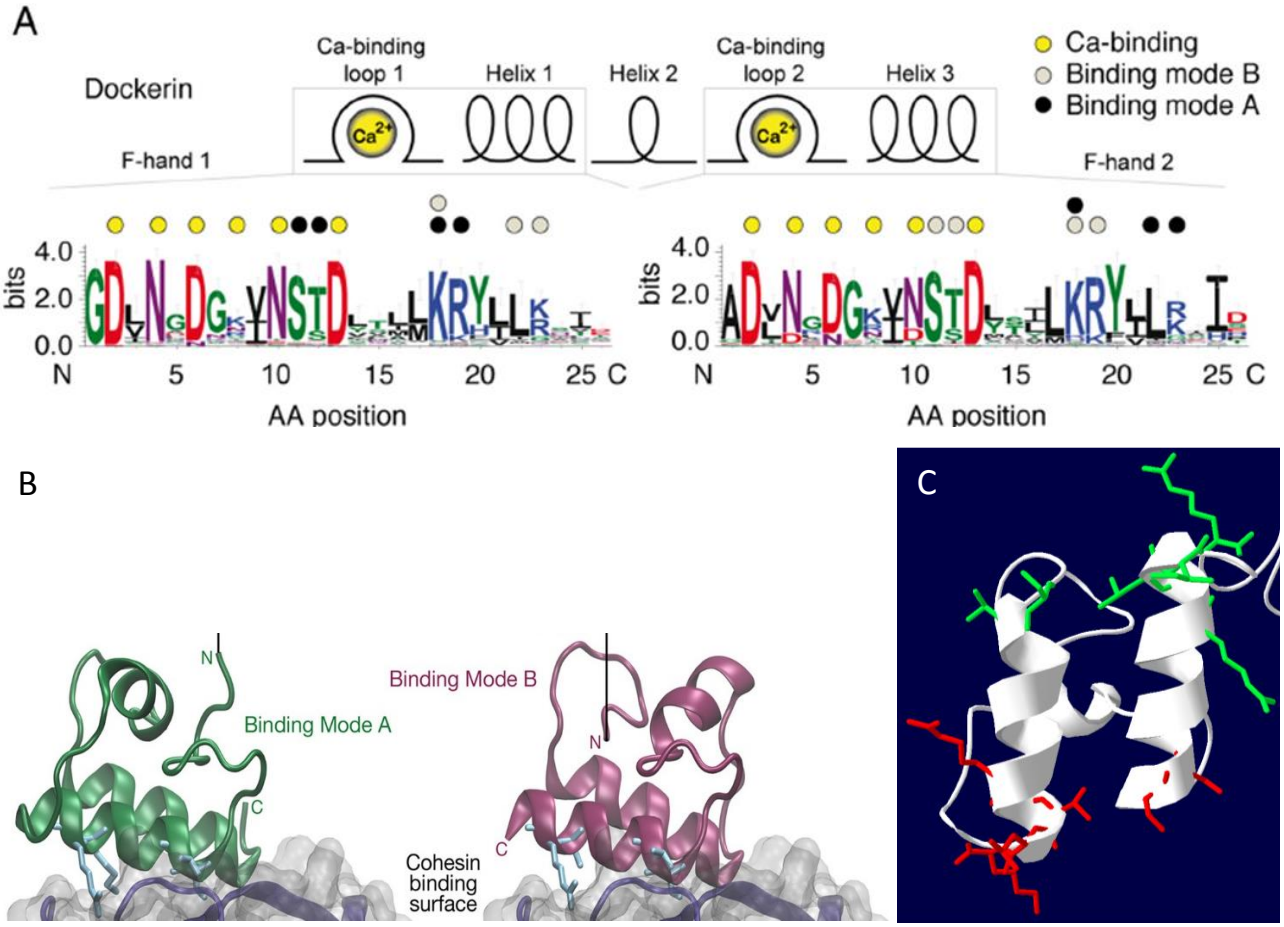


Figure 5. The two binding modes of the dockerin domain. A) The secondary structure, consensus sequence and conservation of the residues of the duplicated sequences of 65 putative *C. thermocellum* type-I dockerins. Depicted above the consensus sequence logo are the calcium binding residues (yellow dots) and the binding residues of the two different binding modes (grey and black dots) of dockerin to cohesin mainly responsible for the high affinity interaction between the two domains. The height of the letter stacks represents the conservation stage of each residue and the relative heights of each letter represent the relative frequency of each residue at that position. In the consensus sequence the residues responsible for the calcium binding and the two binding modes are found at identical position in both of the doubled loop-helix motifs. B) Dockerin interacting with the cohesin domain in each of its binding modes. C) Illustration of the Ct endoglucanase D dockerin domain showing some of the main residues responsible for the interaction with the cohesin domain in the two binding modes, each color (red or green) representing one binding mode. The residues of each of the binding modes show high symmetry with the residues of the other binding mode and are located at the opposite ends of the folded domain. Figures A and B taken from Jobst et al. (2015). Model in Figure C was imaged with Swiss-PDB viewer.

Although the dockerin is a highly conserved domain and some of its residues are essentially crucial to the binding affinity, this is not the case for all of the residues. For example it has been shown that under (normal conditions) at room temperature (RT) the dockerin is able to bind to the cohesin even when it is significantly truncated from either end of the domain (Karpol et al. 2008). There were even notable differences in the effect of the truncation between the two termini of the domain for the interaction, since more than 20 residues could be deleted from one end to still remain the

binding capacity when truncation of already 16 residues from the other end was enough to abolish the interaction completely.

Another, only recently discovered feature of the dockerin is the “intramolecular clasp” formation of the domain (Slutzki et al. 2013, Chen et al. 2014). The clasp is formed between two or more complementary, typically hydrophobic, aromatic or charged residues at each end of the domain (located before the first calcium binding loop and after the second loop-helix motif) binding the ends together (Figure 6)(Slutzki et al. 2013, Chen et al. 2014). The clasp formation is considered to stabilize the overall structure of the dockerin by affecting the folding of the ends of the domain. This contributes also to the overall fold of the domain and the correct orientation of the residues interacting with cohesin thus affecting the recognition and binding affinity of the interaction (Slutzki et al. 2013). Mutations in two of the clasp forming residues in *Ruminococcus flavefaciens* ScaA dockerin domain were accordingly found to reduce its affinity to its cohesin counterpart by approximately an order of magnitude and to reduce the thermostability of the domain by 20˚C (Slutzki et al. 2013).

The interactions between cohesin and dockerin domains are mainly species specific which means that for example the dockerins or the cohesins from *C. thermocellum* do not recognize or bind the proteins corresponding to their counterparts from another species of the *Clostridium* spp. (Pagès et al. 1997). Although the structures of the cohesin domains from different species are very similar and their sequences show about 50% identity or similarity with each other (Pagès et al. 1997), the residues responsible for the recognition and high affinity of the cohesin-dockerin interaction are, however, very different between species. These residues simultaneously show both high conservation within a certain species but high dissimilarity between them (Pagès et al. 1997, Carvalho et al. 2003). This results in the formation of favorable interactions between the recognition residues of the cohesin and dockerin domains of one species because of their similar or compatible nature while between species these interactions are not formed because of the incompatibility between the residues.

The double binding mode of dockerins to cohesins is thought to increase the conformational space or structural flexibility available to the scaffoldin-borne enzymes and thus enable improved substrate recognition and hydrolysis capabilities (Smith and Bayer 2013). It also has been suggested that the duplication of the sequence would simultaneously confer robustness and flexibility to the protein in respect of function altering mutations by allowing some variations to the sequence. The

proteins would be susceptible to mutations even in the conserved regions to allow for the exploration of interspecies recognition while retaining their high affinity for the cohesins of the same species. In this way the cellulosomal systems of the bacteria living in mixed species colonies and communities would be more compatible to each other and allow for a more flexible and efficient use of resources in mixed species bacterial communities (Jobst et al. 2015).

The cohesin-dockerin interaction has been previously utilized for example in synthetic macromolecular scaffolds used for screening for suitable mixtures of enzymes and enzyme activities for commercial enzyme preparations (Hahm et al. 2015) and in substrate channeling of synthetic metabolic pathways *in vivo* (Kim and Hahn 2014). The use of synthetic designer scaffoldins has also been applied on numerous occasions, in studying the roles and functions of the cellulosomal proteins among other multiple applications, which has shown the applicability and mouldability of the system (Artzi et al. 2017). Many more applications for the cellulosomal components have been proposed earlier and the potential of the system for different biotechnical applications and processes has even been described to seem unlimited (Bayer et al. 1994).

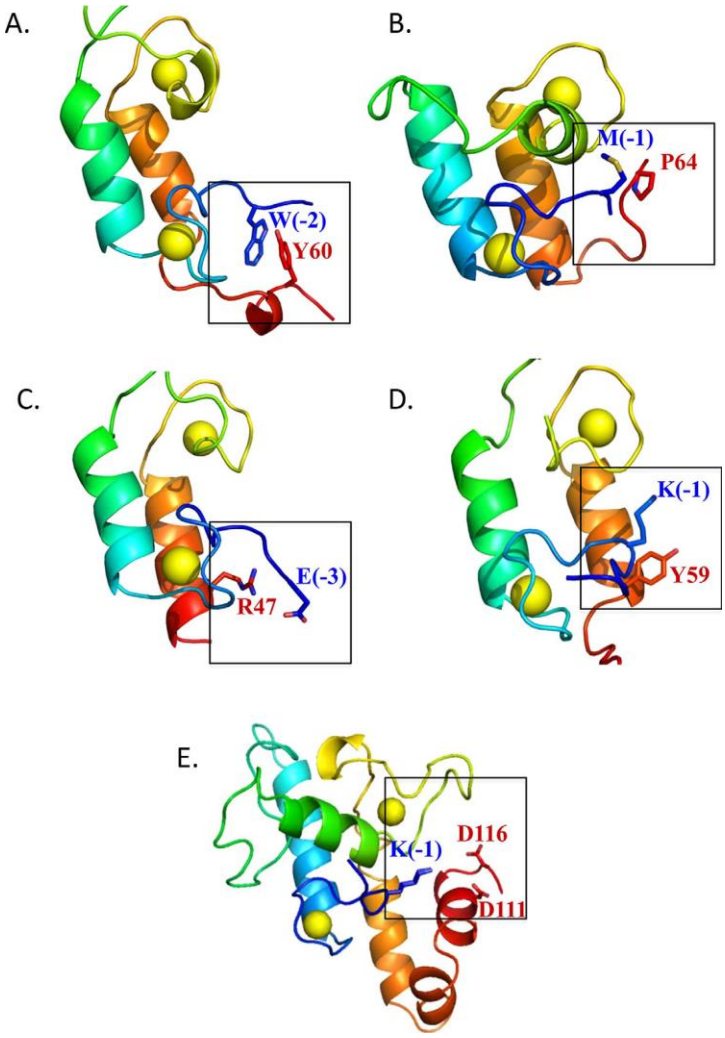


Figure 6. Intramolecular clasps of different dockerin domains. A) *C. thermocellum* type-II dockerin (PDB ID 2B59). B) *C. thermocellum* type-I dockerin (PDB ID 4FL4). C) *C. perfringens* dockerin (PDB ID 2OZN). D) *B. cellulosolvens* dockerin (PDB ID 2Y3N). E) *R. flavefaciens* CttA dockerin (PDB ID 4IU2). Depicted on the figures are dockerin structures of different bacterial species showing the intramolecular clasp formation between (highlighted) residues at both ends of the domain. The negative residue numbering of the residues shown in blue refers to the position of the residues before the glycine residue of the first calcium binding loop of the domain (Figures 5 A and 4 A). The *C. thermocellum* type-I dockerin depicted in Figure B (PDB ID 4FL4) corresponds mostly to the DocD used in this thesis work, although with differences at both ends of the domain rising from the modifications made to the domain in this thesis work (Table 2 and supplementary files). Figure taken from Slutzki et al. (2013).

1.3. Thesis hypothesis, motivation for the study

Screening for new cellulases, particularly for new combinations of catalytic domains and CBMs, is a viable strategy to enhance cellulose hydrolysis as shown earlier at VTT (Carrard et al. 2000, Voutilainen et al. 2014). The different properties of cellulases that have been and are further sought to be enhanced by protein engineering are e.g. lowered end-product inhibition, increased specific activity and higher thermostability (Viikari et al. 2012, Payne et al. 2015, Artzi 2017). Several different strategies have been developed to enhance the enzymes' properties. These methods

include for example directed evolution, random mutagenesis, rational design and domain shuffling approaches (Voutilainen et al. 2014, Payne et al. 2015).

The ”traditional” way for this type of protein engineering in creating new 2-domain enzymes is to design the mutations, clone and transform the DNA constructs and to express, purify and characterize the recombinant enzymes. Since the whole procedure takes some time, a method to combine two protein domains only after producing them separately would significantly reduce the amount of time and work needed for the production of these type of fusion enzymes (Figure 7 and Table 1). For example in order to screen for e.g. every possible combination between N number of different catalytic domains and M different CBMs the amount of constructs needed to be created/ordered, cloned, transformed, produced and purified would be N x M. Then again, if these different domains could be produced separately and combined with each other only after purification the amount of protein constructs would be N+M. As the amount of these different domains increases higher and higher the relative amount of work needed to be done between these different methods decreases all the time in favor of producing the domains separately as can be seen in Table 1.

The dockerin-cohesin interaction has been successfully utilized previously in creating active enzyme complexes that contain the catalytic domain and CBM -counterparts (Carrard et al. 2000, Caspi et al. 2006, Caspi et al. 2008, Vazana et al. 2010). For example in a study by Carrard et al. (2000) the *C. thermocellum* endoglucanase CelD (Ct Cel9A), with its naturally occurring dockerin domain, was combined with different (*C. thermocellum*) cohesin-CBM fusion proteins of different origins. As a result, functional enzyme complexes with varying specific activities and affinities to cellulose were obtained, with the *C. thermocellum* CipA associated CBM (CBM3) as the most effective CBM to enhance the cellulose hydrolysis (Carrard et al. 2000).

In this thesis work the high affinity of the type-I cohesin-dockerin interaction was attempted to be transferred to fungal GH7 cellobiohydrolases in order to facilitate their faster production and screening for enhanced crystalline cellulose hydrolysis. The different dockerin-CBM fusions were planned to be expressed either in bacterial or yeast expression hosts, while the cohesin-catalytic domain fusions would be produced separately in yeast *S. cerevisiae*. Finally the complex formation between the different fusion protein counterparts was to be studied and the complex characterized according to its thermostability and enzyme activity at different temperatures. The results were to be compared to a reference enzyme, in which the catalytic domain and the CBM would be

connected directly to each other with a linker peptide, instead of the cohesin and dockerin domains (Figure 9). This was done in order to see whether the cohesin-dockerin mediated binding should have any effect on the enzymatic activity or thermostability of the enzyme and whether there would be any other differences between the two systems.

The heterologous expression of GH7 family fungal cellobiohydrolases is known to be challenging, apparently due to the many (9-10) disulfide bridges in the CBH catalytic domain and several N-glycosylation sites (Grassick et al. 2004). Moreover, the heterologous production (of soluble) dockerin domains has proved to be difficult and the proteins have e.g. suffered of proteolysis in *E. coli* (Fierobe et al. 1991, Murashima et al. 2001, Carvalho et al. 2003). Since the principal motivation behind this study was to ease the workload and speed up the screening of different protein domain combinations, the ease of production of these protein domains is essential for the applicability of the method desired to be created here. That is why the heterologous production of these proteins in different production hosts, the possible obstacles in these systems and the subsequent solutions to them were of particular interest of this thesis work.

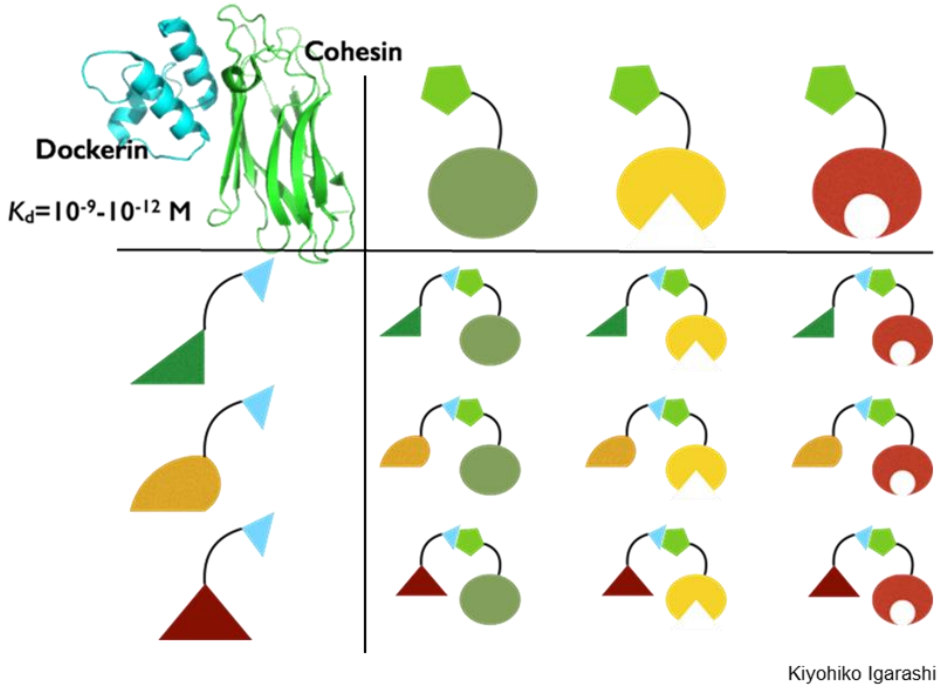


Figure 7. Schematic view of the principle of the method developed in this MSc thesis work for screening of different cellobiohydrolase- CBM combinations with the help of the cohesin-dockerin interaction. Different CBMs (green, yellow and red on the left) would be produced as fusion proteins with dockerin domains (blue triangles) and catalytic domains (green, yellow and red on top) would be produced as fusion proteins with cohesin domains (green pentagons). The different fusion proteins would be connected to each other only after production and purification to yield functional CBM and catalytic domain containing fusion protein complexes (lower right region) connected to each other via the high affinity cohesin-dockerin interaction (upper left region). Figure by Kiyohiko Igarashi, VTT.

Table 1. A comparison between the conventional protein production method and the dockerin-cohesin mediated complex formation method in the production of multidomain cellulases. As the number of different domains needed to be screened for increases, the relative amount of work needed to be done with the dockerin-cohesin -mediated complex formation method decreases when compared to the traditional method, in which every construct has to be produced separately as a fusion protein containing both the catalytic domain and the CBM connected by a peptide linker.

Number of constructs		Number of CBMs					
		1		4		16	
Number of catalytic domains	1	1	2	4	5	16	17
	4	4	5	16	8	64	20
	16	16	17	64	20	256	32

= conventional method
= cohesin-dockerin mediated complex formation

2. Materials and methods

2.1. Strains, genes, plasmids and transformations

2.1.1. *Escherichia coli* expression system

The bacterial production strains used in this project were the *Escherichia coli* BL21 (DE3) and the *E. coli* SHuffle strains T7 and T7 express (New England Biolabs) capable of forming cytoplasmic disulfide-bond containing proteins (Anton et al. 2016) and which are based on *E. coli* B (BL21) strain.

The bacterial gene expression plasmids used in this study (Table 2) were based on the pBAT4 plasmid containing a T7 promoter, an ampicillin resistance selection marker (a beta-lactamase gene) and an origin of replication sequence for amplification in *E. coli* (Peränen et al. 1996).

The genes to be expressed in *E. coli* strains were ordered as synthetic genes or so called G-blocks from Integrated DNA Technologies (USA). The *E. coli* codon optimized fusion genes consisted of different *Clostridium thermocellum* (Ct) dockerin domains (Ct Endoglucanase D dockerin (DocD) or Ct CelS dockerin (DocS)) connected to a cellulose-binding domain (CBM) of either fungal (*Trichoderma reesei* (Tr) cellobiohydrolase I (Cel7A) CBM (CBM1)) or bacterial origin (Ct scaffoldin (CipA) CBM, CBM3) via a varying flexible linker region. Endoglucanase D dockerin (DocD) was fused with the CBM1 (Table 2, Figure 8 A) and the CelS dockerin with the CBM3 with varying linker regions (Table 2, Figure 8 B). The DocS-CBM3 constructs also contained histidine-tags in either end of the fusion protein, to facilitate their affinity purification on a nickel-sepharose column. The different dockerin-CBM -constructs are listed in Table 2 and in more detail in supplementary files. The

synthetic gene fragments included about 60 bp long regions on both 5’ and 3’ end overlapping with the vector to enable cloning into the pBAT4 vector linearized with NcoI and XhoI using the Gibson Assembly Master mix (New England Biolabs, USA) according to manufacturer’s instructions.

E. coli XL-1 blue strain was used as the cloning host. The Gibson assembly mixture was transformed into chemically competent *E. coli* XL-1 blue cells by heat shock method for amplification of the plasmids. After transformation the cells were plated on 100 µg/ml ampicillin containing Luria-Bertani -media (LB-amp, containing 0.5% yeast extract (w/v), 1% tryptone (w/v) and 1% NaCl (w/v)) selection plates (with 1.5% (w/v) agar), from which several colonies were picked to be cultured and grown on LB-amp for plasmid purification. Plasmid purification was done with a QIAprep® Spin Miniprep Kit (Qiagen, Germany) according to manufacturer’s instructions. Plasmids prepared this way were send to sequencing to Gatc-biotech (Germany) to verify their correct sequences. Plasmids were then transformed to chemically competent *E. coli* production strain cells by heat shock method and the transformed cells were plated onto LB-amp selection plates.

Table 2. Fusion protein gene constructs. DocD-CBM1 fusions were expressed in many different *E. coli* expression hosts as well as in *S. cerevisiae* H1910. The DocS-CBM3 fusions were expressed only in *E. coli* BL21 (DE3). In addition to dockerin-CBM fusions, a *S. cerevisiae* produced catalytic domain-cohesin fusion protein was used in this thesis to create the active enzyme complex.

Fusion protein construct	Linker	Mutations	Signal sequence	Expression vector	Plasmid	Expression strain	References/other
Ct Endoglucanase D dockerin (short) + linker 1 + CBM1 (TrCel7A)	TTRRPATTTGSSPGP = linker 1	6 amino acids shorter than the longer dockerin (N-terminal VLY and C-terminal RVI amino acids removed)		pBAT4	FDP5	*E. coli* Shuffle T7, Shuffle T7 express, BL21 (DE3)	Sanni Voutilainen, unpublished work. Supplementary file nnvvnv
Ct Endoglucanase D dockerin + linker 1 + CBM1 (TrCel7A)	TTRRPATTTGSSPGP = linker 1			pBAT4	FDP6	*E. coli* Shuffle T7, Shuffle T7 express, BL21 (DE3)	Linker 1 = partial *Trichoderma reesei CBHI* linker (GenBank ID P62694.1)
Ct endoglucanase D dockerin (short) + linker 1 + CBM1 (TrCel7A)	TTRRPATTTGSSPGP = linker 1	2 x N-glycosylation sites removed from dockerin*. 6 amino acids shorter than the longer dockerin**	Tr XYN2 secretion signal	pSVEmptyENO	FDP12	S. cerevisiae H1910	*NST and NSS to NSA in both of the calcium binding loops. **N-terminal VLY and C-terminal RVI amino acids removed
Ct Endoglucanase D dockerin + linker 1+ CBM1 (TrCel7A)	TTRRPATTTGSSPGP = linker 1	2 x N-glycosylation sites removed from dockerin*	Yeast alpha factor pre-sequence	pSVEmptyENO	FDP13	S. cerevisiae H1910	*NST and NSS to NSA in both of the calcium binding loops
Te Cel7A+ linker 6 + Ct CipA cohesin2	GNPPGGNRGTTGNPPGG NRGTT = linker 6	2x N- glycosylation sites removed from catalytic domain	TeCel7A	pSVEmptyENO	FDP14	S. cerevisiae H1910	Linker 6 = partial Trichoderma reesei CBHI linker (GenBank ID P62694.1) Sanni Voutilainen, Unpublished work
6His + Ct CelS dockerin + linker 1 + Ct CipA CBM3 (construct 1)	TTRRPATTTGSSPGP = linker 1			pBAT4	FDP21	*E. coli* BL21 (DE3)	
Ct CelS dockerin + linker 2 + Ct CipA CBM3 + 6 His (construct 2)	TTEPATPTTPVTTPTTT= linker 2			pBAT4	FDP22	*E. coli* BL21 (DE3)	Linker 2 from GenBank accession number AF283514 (Carrard et al. 2000)
Ct CelS dockerin + linker 3 + Ct CipA CBM3 + 6 His (construct 3)	PGAASSSSGS = linker 3			pBAT4	FDP23	*E. coli* BL21 (DE3)	Linker 3 from GenBank accession number AF283515 (Carrard et al. 2000)
Ct CelS dockerin + linker 4 + Ct CipA CBM3 + 6 His (construct 4)	GGSVVPSTQPVTTPPA = linker 4			pBAT4	FDP24	*E. coli* BL21 (DE3)	Linker 4 from GenBank accession number AF283517 (Carrard et al. 2000)
Ct CelS dockerin + linker 5 + Ct CipA CBM3 + 6 His (construct 5)	PTPTPTTPTPTPTTPTPTPT = linker 5			pBAT4	FDP25	*E. coli* BL21 (DE3)	Linker 5 from *Cellulomonas fimi* Cex (GH10 xylanase)(Sanni Voutilainen, personal communication)

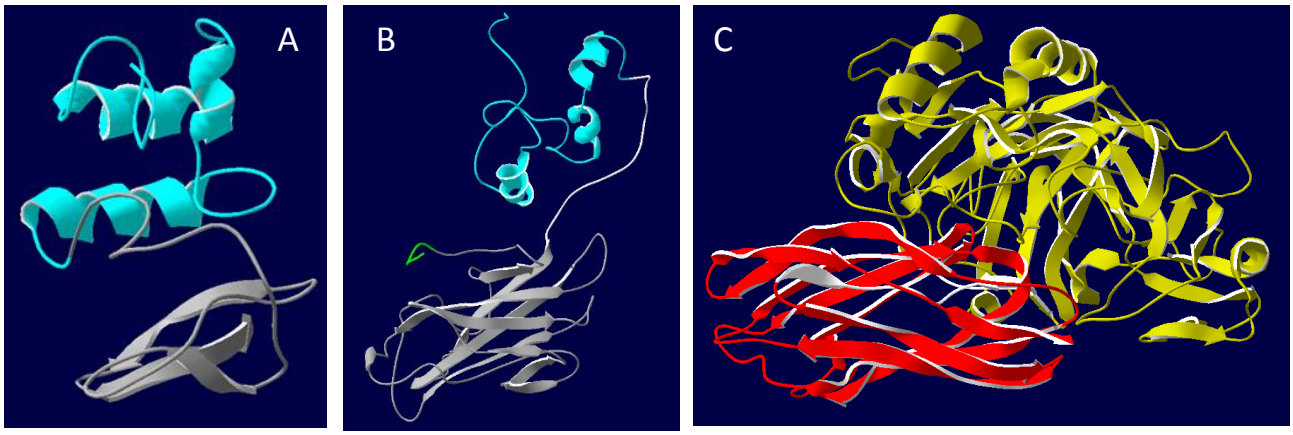


Figure 8. Dockerin and cohesin fusion protein constructs. A) A model of a DocD-CBM1 fusion protein. B) A model of DocS-CBM3 fusion protein. The CBM parts of the proteins are shown in grey and dockerins on light blue. Histidine-tag located at the C-terminus of DocS-CBM3 fusion is shown in green. The size of the DocS-CBM3s, consisting of some ~240 amino acids is much greater than that of the DocD-CBM1s which consists of only <120 amino acids. C) A model of the TeCel7A-cohesin fusion protein. The cohesin domain consisting of only β-strands is depicted on red and the catalytic domain on yellow. Figures depicted here are not exact real structures of the proteins and are meant to be shown here just as an illustration. Images of the models are made with Swiss PDB-viewer.

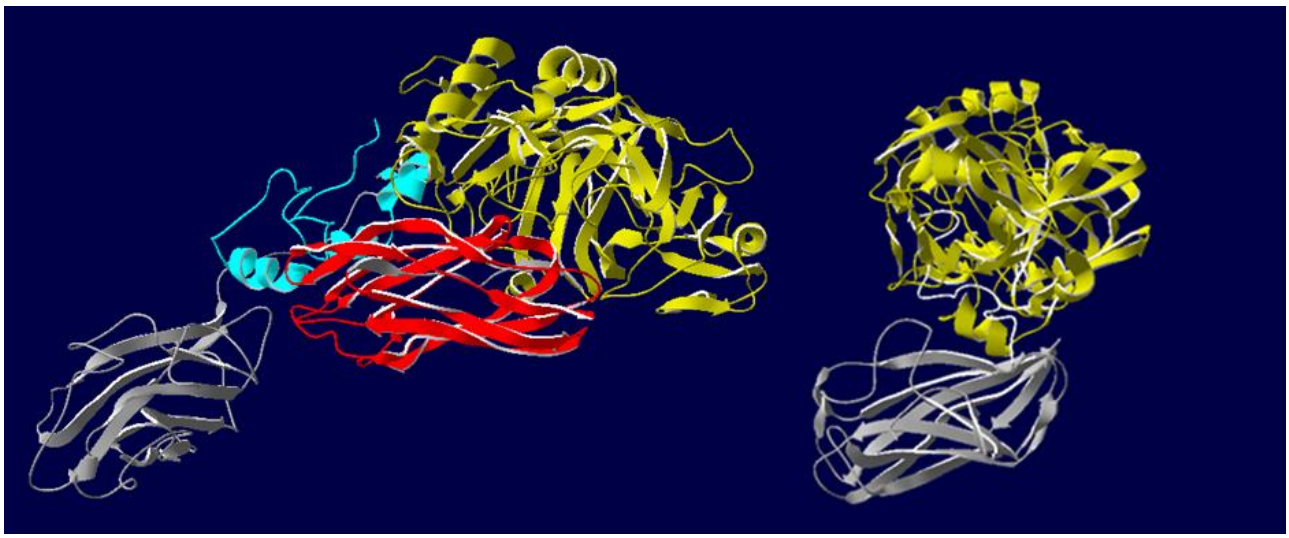


Figure 9. On the left side of the figure is an illustration of the complex between the DocS-CBM3 and TeCel7A-cohesin fusion proteins. On the right side of the figure is an illustration of the directly linked TeCel7A-CBM3 fusion protein used as a control protein for the enzyme complex in this thesis work. The CBM3s of the proteins are depicted on grey, the catalytic domains on yellow, dockerin on light blue and cohesin on red. From the figure it can be seen that the enzyme complex with the dockerin-cohesin –linkage in between is a far larger protein than the directly linked TeCel7A-CBM3 fusion. The two proteins and their corresponding domains are not oriented in a similar manner and may not be in exact correct scale. Also the structures of the models are not exact real structures of the proteins and are meant to be shown here just as an illustration. Images of the models are made with Swiss PDB-viewer.

2.1.2. *Saccharomyces cerevisiae* expression system

Expression of DocD-CBM1 fusion constructs was tested, in addition to *E. coli,* also in *Saccharomyces cerevisiae* H1910. The DocD –CBM1 gene fusions were codon optimized for *S. cerevisiae* expression and synthesized by Integrated DNA Technologies. The two putative N-glycosylation sites in each of

the dockerins were removed by mutating NST and NSS to NSA. Two different signal sequences and two variants of the DocD were tested (see Table 2 for details). Besides the N-glycosylation site mutations, the gene constructs coded for the same mature proteins as the dockerin-CBM1 -constructs to be produced in *E. coli*. The constructs were cloned in *S. cerevisiae* expression vector (pSVEmptyENO)(Voutilainen et al. 2014) containing a constitutive Sc ENO1 promoter, an URA3 selection marker for uracil selection in *S. cerevisiae*, an ampicillin selection marker gene for selection in *E. coli* and origin of replication sequences for *S. cerevisiae* and *E. coli.*

The yeast cells were transformed with *Eco*RI and *Xho*I linearized pSVEmptyENO vector and either of the inserts containing different version of DocD-CBM1, single stranded carrier DNA (salmon sperm), polyethylene glycol and lithium acetate according to Gietz and Woods (2002), and plated onto Y-min SCD-ura -selection plates. The inserts contained about 60 bp overhangs on both 5' and 3' end to allow cloning with homologous recombination. The Y-min SCD-ura -medium contained 2% (w/v) glucose; 0.67% sterile filtered yeast nitrogen base w/o amino acids and 5% (v/v) synthetic complete stock for yeast minus uracil for selection. The plasmids were isolated from the yeast and propagated in *E. coli* XL-1 blue for sequencing as described before.

In addition to the dockerin-CBM proteins, a catalytic domain-cohesin fusion counterpart (Figure 8 C) was needed in order to produce an active enzyme through complex formation between cohesin and dockerin domains (Figure 9). The catalytic domain-cohesin fusion counterpart, was produced in *S. cerevisiae* H1910 similarly to the yeast production described earlier (Sanni Voutilainen, unpublished work). The cohesin-catalytic domain fusion protein consisted of *Talaromyces emersonii* (Te) Cel7A cellobiohydrolase core domain, a linker region and a Ct CipA cohesin2 (cohesin) domain (TeCel7A-cohesin, Table 2, Figure 8 C supplementary file 3).

2.2. Heterologous production of the proteins

2.2.1. E. coli

E. coli strains were cultivated in LB-amp -media supplemented with 5 mM $CaCl_2$ in Erlenmeyer shake flasks in different temperatures for different periods of time to test their effect on the production levels and correct folding of the target proteins. The different sized shake flasks were always more than 20 times the volume of the final liquid volume in them and they were constantly shaken at around 240 rpm. The pre-cultivation was done at 37 ˚C for the *E. coli* BL21 (DE3) strains and at 30 ˚C

for the *E. coli* SHuffle strains according to manufacturer's instructions, since the SHuffle strains were more sensitive to high temperatures. The cultivation was continued until the cells had reached the logarithmic growth phase (OD600s were ~0.1 to 0.8), after which the expression of the target protein was induced and the cultures were put to desired growth temperatures. Induction of expression was done with a final concentration of 0.4 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG). The tested expression conditions for the *E. coli* strains were 3 hours or overnight (O/N) at 37 ˚C, O/N at 30 ˚C and 3 days at 16 ˚C. After the desired cultivation time the cells were harvested by centrifugation in 15 ml conical centrifuge tubes at 3200 rcf for 10 minutes in Eppendorf Centrifuge 5810 R (Eppendorf AG, Germany), after which the supernatants were discarded and the pellets were frozen and stored at -20 ˚C.

2.2.2. S. cerevisiae

S. cerevisiae was cultivated in Y-min SCD-ura -media supplemented with 5 mM $CaCl_2$ at 30 ˚C, 240 rpm in different sized shake flasks, so that the shake flasks were at least 10 times the size of the liquid volume guaranteeing sufficient oxygen levels. After 3 days the cultures were taken out of the incubator, centrifuged for 10 minutes in either 50 ml conical centrifuge tubes at 3220 rcf in Eppendorf Centrifuge 5810 R or at 5000 rcf with a Sorvall LYNX 4000 centrifuge (Thermo Scientific) and the supernatants were salvaged and stored in -20 ˚C for further studies.

The effect of buffering of the cultivation medium on the production levels of the proteins was studied by cultivating the yeast also in buffered Y-min SCD-ura (Y-min SCD-ura pH 6). Y-min SCD-ura pH 6 was prepared in the same way as Y-min SCD-ura, but, instead of water, to a solution containing 20 g/l of succinic acid and 12g/l of NaOH and the pH had been finally adjusted to 6 by addition of 5 M NaOH and finally supplemented with 5 mM $CaCl_2$. The cultivation in buffered solution proceeded similarly to the unbuffered media cultivation and finally the pHs of the two cultivation supernatants from approximately the same phase and time point of the cultivation were analyzed and compared to each other.

2.3. Purification of the proteins

2.3.1. Protein extraction, SDS-PAGE and Western blot analysis

To extract and examine the intracellularly produced proteins in *E. coli* cultivations, the frozen cells were thawed and lysed. This was done by resuspending and incubating the cells in lysis solution 1 for an hour in room temperature (RT). Lysis solution 1 contained B-PER® Bacterial Protein Extraction Reagent (product #78248, Thermo Scientific, USA), cOmplete, EDTA-free protease inhibitor

(11873580001, Roche Diagnostics GmbH, Germany) in from one up to five times concentration of manufacturer's instructions, Lyzozyme from egg white (62971, Sigma-Aldrich Chemie GmbH, Belgium) and DNase I from bovine pancreas (10104159001, Roche Diagnostics GmbH, Germany). After cell lysis the soluble fraction of the lysate was separated from the insoluble by centrifuging at 21000 rcf for 10 minutes.

Alternatively, cell lysis was also done by sonicating the frozen and thawed cells resuspended in 50 mM Tris-HCl -buffer (pH 7.5) containing lysozyme, DNase and protease inhibitor. Sonication was done with Sonics Vibra Cell™ VCX500 sonicator (Sonics & Materials, Inc., USA) with 5 x 15 seconds sonication with 45 seconds in between with an amplitude of 40% on ice. Cell lysis by sonication in Tris-HCl -buffer was done in order to see whether any of the unknown components of the lysis solution 1 (B-PER reagent) had any effect on the complex formation or the cellulose affinity purification of the proteins.

To analyze the target protein content in the cell lysate, samples of the soluble and insoluble fractions were heat denatured in the presence of mercaptoethanol and analyzed with SDS-PAGE using Criterion TGX Stain-Free Precast 4 –20% gradient gels (Bio-Rad, USA). The proteins were either imaged on the gel with Criterion Stain Free Imager (Bio-Rad, USA), stained with PageBlue™ Protein Staining Solution (Thermo Scientific) or blotted from the gel to a 0.2 µm nitrocellulose membrane using a Trans-Blot Turbo Transfer Pack (Bio-Rad, USA) and a Trans-Blot Turbo transfer system with a constant current of 2.5 A for 7 minutes.

Detection of the target protein on the membrane was done by Western blotting (immunoblotting). After blocking with 2% milk, the membrane was incubated with either anti-CBM1 -monoclonal antibody or the anti-histidine-tag -monoclonal antibody (anti-His-tag –antibody), then with a secondary antibody-alkaline phosphatase conjugate (Goat Anti-Mouse IgG, (H+L) -AP Conjugate, 1 ml, Bio-Rad, USA) and finally by color detection with BCIP/NBT color development substrate (ref. 3771, Promega, USA).

2.3.2. Affinity purification on cellulose material

The purification of the target proteins, which did not contain a His-tag, was tested with affinity purification in a cellulose (Avicel) column in the presence of 1 M ammonium sulfate (Sugimoto et al. 2012) or by a batch type of purification in conical 50 ml centrifuge tubes in similar conditions for some of the yeast cultivation supernatants.

To prepare the Avicel column, Avicel 101 (Fluka) was rinsed sufficiently with distilled water to remove the fine particles that could clog the column and slow the purification process. The rinsing was done by mixing the Avicel with water, allowing it to settle down on the bottom of a decanter glass for the most parts and then by discarding the supernatant and repeating these steps again. The rinsed Avicel was packed into a column and equilibrated with 1 M ammonium sulfate.

To separate and purify the target proteins from the yeast culture supernatant or the bacterial cell lysates, these were made 1 M with ammonium sulfate before applying them into the Avicel column. Because of aggregate formation in the bacterial cell lysates at this point, the cell lysates were centrifuged for 5-10 min at 3220 rcf and the pellet was discarded. The yeast and bacterial supernatants were then ran through the affinity column and the column was washed with 12 column volumes (CV) of 1 M ammonium sulfate. After washing the target protein was eluted from the column with double distilled water containing 5 mM of $CaCl_2$, and the elution fractions were gathered and analyzed on an SDS-PAGE gel and/or by Western blotting. The degree of purification and elution efficiency was also analyzed from the starting sample, flowthrough and wash fractions and from the Avicel in the column, by incubating the liquid or Avicel samples with SDS-PAGE loading dye containing mercaptoethanol at ≥ 90 ˚C for 10 minutes and analyzing them on SDS-PAGE and by a Western blot. Because of the aggregation the purification was also tried with lower ammonium sulfate concentrations (0.5 M and 0.25 M) to test and to try to prevent the precipitation effect of the ammonium sulfate in the purification.

Batch purification of the yeast supernatants with Avicel was done in 50 ml conical centrifuge tubes in a head over head rotator in varying temperatures. All the preparations of the supernatant before mixing with Avicel were done in a similar way to those of the Avicel column purification preparations, except for that the Avicel was not washed prior to purification and the supernatant, Avicel and washing solutions were cooled down to 4 ˚C prior mixing with each other. Avicel was mixed to yeast supernatant in 20 mg/ml ratio. Mixing the supernatant and binding of proteins to Avicel was done at 4 ˚C in a head over head mixer for an hour. The mixture was centrifuged for 10 minutes at 3220 rcf and the supernatant was discarded. The pellet was resuspended in 25 ml of cold 1 M ammonium sulfate supplemented with 5 mM $CaCl_2$ at RT and washed by vigorously vortexing. This was repeated two times while in the middle centrifuging and discarding the supernatant. After

washing, the proteins were eluted by resuspending and incubating the Avicel in 5 ml of water supplemented with 5 mM $CaCl_2$ in a head over head rotator at RT for an hour. Also an additional elution for O/N with fresh elution solution was made.

The samples from different purification steps were concentrated up to 15 times with Vivaspin 4 Turbo 5000 MWCO membranes (Sartorius), meanwhile changing the buffer to 50 mM Tris-HCl (pH 7) supplemented with 5 mM $CaCl_2$. Concentration and buffer change was done in order to ease the detection of the target proteins in the samples. To see whether the proteins were still bound to the Avicel after elution, a sample of the Avicel was resuspended in a sufficient volume of water to match the original concentrations of the other samples, incubated with SDS-PAGE loading dye at ≥ 90 ˚C for 10 minutes to denature the bound proteins after which the supernatant was salvaged for analysis. Finally the ready samples were analyzed by SDS-PAGE and Western blotting.

2.3.3. Histidine-tag based affinity purification

Protein constructs which contained a His-tag were purified from the soluble fractions of cell lysates by affinity purification in a HisTrap FF crude 1 ml (GE Healthcare) nickel-sepharose column using ÄKTA purifier (Amersham biosciences). Two phosphate buffers (containing 20 mM sodium phosphate, 500 mM NaCl and imidazole, pH 7.4) were made with different imidazole concentrations (10 mM (purification buffer A) and 500 mM (purification buffer B)) to be used in the purification during different phases. The column was equilibrated with purification buffer A.

The soluble fraction of the cell lysate was diluted 1:1 with purification buffer A, so that the binding of the proteins to the column was done in the presence of 5 mM imidazole, to reduce unspecific binding to the column. The diluted cell lysate supernatant was centrifuged at 3220 rcf, to remove the precipitate formed after admixing of the purification buffer. The supernatant was loaded into the column at a rate of 1 ml/minute, after which the column was washed with purification buffer A for 5 column volumes. Samples of both the flowthrough and wash fractions were gathered. The bound proteins were eluted with 12-20 CV linear gradient from 0 to 100% of buffer B. The elute from the gradient was collected as 500 µl fractions to be later analyzed for protein content in SDS-PAGE and to be pooled and gathered as the purified proteins. The buffer of the pooled fractions was changed to 4 ml of 50 mM Tris-HCl buffer (pH 7) supplemented with 5 mM $CaCl_2$ in Econo-Pac 10DG Desalting Columns (732-2010, Bio-Rad Laboratories Inc., USA) after which the proteins were stored at -20 ˚C.

2.4. Characterization of the proteins

2.4.1. Storage stability of the fusion proteins

The stability of the His-tag purified *E. coli* produced DocS-CBM3 fusion proteins was tested by storing samples of the proteins in different temperatures and comparing them to freshly purified proteins. The protein samples were stored at 4 ˚C and -20 ˚C for 19 or 22 days after which they were analyzed on SDS-PAGE.

2.4.2. Cohesin-dockerin complex formation

The complex formation was done by mixing the proteins in suitable proportions according to their molar concentrations in a microcentrifuge tube by gently pipetting the mixture up and down. Determination of the protein concentration was made by measuring the purified proteins' absorbance at 280 nm in a 1 cm path length acrylic cuvette (Sarstedt, Germany) with an Ultrospec 2100 pro -spectrophotometer (Amersham Biosciences). The absorbance values were then divided with the proteins' theoretical extinction coefficiency values computed by the ProtParam online protein parameter computation tool (Gasteiger et al. 2005). The proteins were diluted to their desired concentrations with 50 mM sodium acetate buffer (pH 5) supplemented with 5 mM $CaCl_2$ (hydrolysis buffer). The mixes were incubated statically at RT and at 4˚C for different time periods ranging from 10 minutes to O/N. The mixes were then analyzed on native SDS-PAGE without heat denaturation or mercaptoethanol in the loading buffer using Criterion TGX Stain-Free Precast 4 – 20% gradient gels and by running the gel with varying voltages from 180 V to 240 V. The proteins in the gels were visualized by using Criterion Stain Free Imager.

2.4.3. Soluble substrate hydrolysis

4-methylumbelliferyl-β-D-lactoside (4-Methylumbelliferyl β-D-lactopyranoside, MULac) that was used here as the hydrolysis substrate, is a soluble substrate that can be used to measure the activities of cellobiohydrolases. Cellobiohydrolases are able to hydrolyse MULac so that a fluorescent molecule, metylumbelliferyl (MU), is released and can be detected by a fluorescent measurement. The activity of the TeCel7A-cohesin fusion was measured in this experiment alone, without the dockerin-CBM counterpart, since the method only measures the activity of the catalytic domain. Also the activity of the control enzymes, the Sc produced TeCel7A core preparation (Voutilainen et al. 2010) and the Sc produced TeCel7A-CBM3 fusion preparation (Voutilainen et al. 2014), was measured.

For each of the measurements a final reaction mixture of 100 µl with a 2 mM MULac concentration and a 0.2 µM enzyme concentration in 50 mM NaAc -buffer (pH 5) was made in a single well of a black flat-bottomed 96 well microtitre plate. The reaction mixture also contained approximately 6.25% (v/v) DMSO that had been used for the dissolving of the MULac in the substrate stock solution. The reactions were started by the addition of the substrate to the solution and stopped after 2.5; 5; 7.5 or 10 minutes by the addition of 100 µl of 1 M Na_2CO_3. The enzyme zeros were done otherwise in the same way as the other samples but the addition of enzymes was done only after the addition of the 1 M Na_2CO_3. Standard curve reactions were made by mixing 100 µl of 2 mM MULac in 50mM NaAc -buffer (pH 5) with 100 µl of from 3 to 48 µM MU in 1 M Na_2CO_3. All of the measurements and time points were performed as triplicates. The detection of the released MU was done by measuring the fluorescence emitted from each well at 460 nm after excitation at 355 nm, with Varioskan 3.01.15 microplate reader (Varioskan, Thermo electron corporation, USA) and results were calculated from the cellobiose standard curve after reduction of the background.

2.4.4. Thermostability

Thermostability of the proteins and the protein complex was studied with circular dichroism (CD). The proteins' CD spectra at different temperatures were measured with Chirascan CD Spectrometer (Applied Photophysics, United Kingdom), using QS High Precision Cell 1 mm light path quatz SUPRASIL cuvette (100-1-40, Hellma Analytics, Germany) and TC125 temperature control heater (Quantum Northwest, USA) and AWC100 recirculating cooler (Julabo, Germany) to control the sample temperature. The spectra were recorded from 190 nm to 250 nm in temperatures ranging from approximately 20 ˚C (or RT) to ~85 ˚C. Spectra of the proteins at specific temperatures were obtained from the averages of two parallel measurements at the same or very close (within a degree) temperature of each other. The measurements were performed in 10 mM NaAc, pH 5 supplemented with 5 mM $CaCl_2$ or in 10 mM Tris-HCl pH 7 supplemented with 5 mM $CaCl_2$ using 3 µM protein concentration. Additionally the CD spectra of DocS-CBM3 was also measured in 10 mM NaAc-buffer (pH 5) without additionally supplemented $CaCl_2$ and with or without the presence of 300 µM EDTA (ethylenediaminetetraacetic acid).

When measuring the spectra with temperature ramping mode from RT to 85 ˚C with 2 ˚C measuring steps, however, the measuring was made with one measurement per step with 0.5 second measurement for each measured wavelength and a 0.5 nm distance in wavelength. Since with temperature ramping the temperature in the sample lagged a few degrees behind the set

temperature of the heater, the temperature in the sample continued to rise during each measurement. Thus the actual temperature of the sample within each sample spectra measurement at each temperature point varied with approximately 1.1 ˚C between the initiation and the end of the measurement. The actual temperature of the sample was monitored the whole time with a temperature probe and data temperatures are presented as the temperature of the actual sample.

2.4.5. Cellulose hydrolysis

The complex formation between the TeCel7A-cohesin and the DocS-CBM3 was done so that the final concentration of the mix was 2 µM for the TeCel7A-cohesin and 4 µM for the DocS-CBM3. The mixes were incubated O/N statically at 4˚C, to allow for complex formation to happen as done by Carrard et al. (2000).

To make the reaction mixture, 162 µls of 2% (w/v) stock mixture of Avicel Ph105 in hydrolysis buffer (50 mM sodium acetate buffer (pH 5) supplemented with 5 mM $CaCl_2$) was pipetted into a microcentrifuge tube with 71 µls of the same buffer. The substrate stock solution was constantly stirred with a magnetic stirrer at RT prior to and during the making of the reaction mix to ensure an even amount of substrate in every reaction tube. The reaction tubes were put to a Eppendorf Thermomixer comfort dry block heating and cooling shaker (Eppendorf AG, Germany) set to 50 ˚C, 60 ˚C or 70 ˚C and 1400 rpm to incubate and warm up to reaction temperature before initiation of the reaction. The reactions were started by the addition of 91 µls of the diluted enzyme solutions and incubated for 3 to 24 hours at the set temperature and at 1400 rpm. The final reaction mixtures consisted of 1% (w/v) Avicel and ~0.56 µM enzyme solution in hydrolysis buffer. The reactions were stopped at desired time points (0; 3; 5; 16/17.5 or 24 hours) by the addition of 163 µls of 0.5 M NaOH into the mix, after which the tubes were vortexed and put on ice. The tubes were then centrifuged at 20800 rcf with Eppendorf Centrifuge 5417R (Eppendorf AG, Germany) for 5 minutes to pellet down the remaining Avicel after which the supernatant was salvaged and put to freezer. Enzyme zeros (E_0) were prepared and incubated for 3 hours in the same way, except for that the enzyme solutions were put to the mix only after stopping of the reaction. Substrate zeros were prepared and incubated in the same way as other reactions, but instead of enzyme solution, an equal amount of buffer solution was added to the mix. The reactions for each time points were done as triplicates.

The enzyme activities during the reactions were studied by determining the amount of released reducing sugars in the supernatant by the para-hydroxybenzoic acid hydrazine (PAHBAH) method

(Lever 1972) and by comparing them to a cellobiose standard. To determine the amount of released sugars in the reactions, samples from different time points were first diluted to one tenth of the original concentrations in dilution solution (2 parts hydrolysis buffer, 1 part 0.5 M NaOH) and then mixed in 1:1 volume with PAHBAH reagent (1.5% (w/v) PAHBAH in 0.5 M NaOH). The mixes where incubated for 10 minutes at ~95 ˚C, cooled on ice and spinned down. Aliquots of 200 µls of each sample were then put on a microtitre plate and the absorbance was measured at 405 nm with Varioskan. Final results were calculated from the cellobiose standard curve.

2.5. Modeling

The images of the models shown in this thesis as an illustration were visualized using the Swiss PDB-viewer 4.1.0 (Guex and Peitsch, 1997). The models were made using either SWISS-MODEL protein structure-homology modelling server (Arnold et al. 2006) and/or the Phyre2 online protein modelling, prediction and analysis tool (Phyre2 tool, Kelley et al. 2015) used on intensive modelling mode. The models acquired via SWISS-MODEL were based either on different X-ray crystallography or solution NMR structures from Protein Data Bank (PDB) (https://www.rcsb.org/, Berman et al. 2000). The dockerin parts of the DocD-CBM1 -models were modeled using the X-ray diffraction structure of PDB ID 4FL4 from protein data bank (Currie et al. 2012) as a template to which specific modifications and additions were made in Swiss PDB-viewer in order to match the whole length and modifications made in different constructs of the fusion proteins. The CBM1 part of the fusion was modeled using as template the solution NMR structure of PDB ID 1CBH (Kraulis et al. 1989) as a template. The linker of the constructs was modeled by using the Phyre2 tool on intensive mode using the whole sequence of the fusion protein in the initial modeling and then by excluding the dockerin and CBM domains in Swiss-PDB viewer. Finally the different parts were merged into one in Swiss-PDB -viewer.

The DocS-CBM3 fusion, the TeCel7A-cohesin fusion and the directly linked TeCel7A-CBM3 fusion proteins were modeled using only the Phyre2 tool on intensive modelling mode. The model of the complex of TeCel7A-cohesin/DocS-CBM3 was made using the modeled structures of the two counterparts obtained with the Phyre2 tool and by using the X-ray diffraction structure of the cohesin-dockerin complex of PDB ID 4FL4 as a template for the cohesin-dockerin interaction when fitting the two parts of the complex together.

3. Results

3.1. Production and purification of the proteins

3.1.1. DocD-CBM1 fusion proteins

3.1.1.1. E. coli produced proteins

The longer DocD-CBM1 fusion was expressed in *E. coli* SHuffle T7 express both in insoluble and soluble form as can be seen from SDS-PAGE and Western blot in Figures 10 A and 10 B, although the protein was mainly found in the insoluble fraction of the cell lysates. The optimum expression time for the SHuffle T7 express strain at 16 ˚C would seem to be 2 days after induction, after which the amount of protein did not seem to increase or even decreased according to results in Figure 10 B. From the insoluble fractions of the lysates and from the blot it can be seen that the proteins are expressed in the production strains and the proteins can be found around at the level of the 15 kDa marker protein band. This corresponds well to the theoretical molecular weight (TMW) of the proteins which were approximately 12 kDa, varying to some extent with the different versions or constructs of the proteins (Table 2, supplementary files). The amount of protein found in the BL21 (DE3) strains insoluble fractions was much greater than those of the SHuffle T7 express strains', however not much of it was found in the soluble fractions (data not shown). The shorter form of the DocD-CBM1 fusion protein was not expressed as well in *E. coli* as the longer version (Sanni Voutilainen, unpublished work), which is why more efforts were focused on the production of the longer version of the protein in *E. coli* in this thesis.

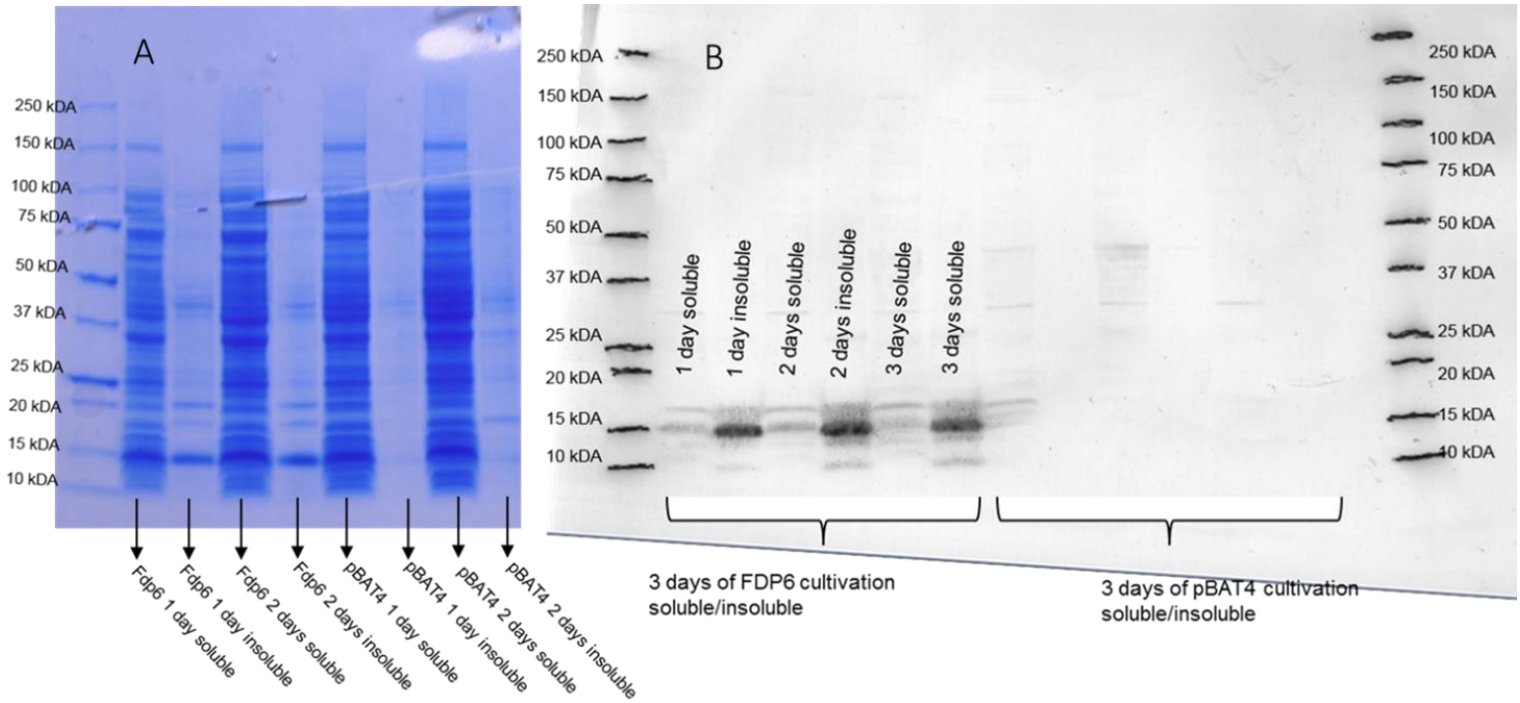


Figure 10. SDS-PAGE image of the DocD-CBM1 fusion protein expression trials in different *Escherichia coli* strains. A) Coomassie stained gel of *E. coli* SHuffle T7 express cells DocD-CBM1 (longer version, FDP6) protein production compared to an empty plasmid control (pBAT4). B) Target protein production of *E. coli* SHuffle T7 express cultivated at 16˚C after induction for different lengths of time. Western blot with anti-CBM1 monoclonal antibody. Proteins are expressed and the yield seems to be strongest in two-day -cultivations. Strong bands are visible in both Figures A and B in 2 days insoluble fractions in the level of 15 kDa marker. The soluble fraction however does not contain much of the protein and the proteins are mainly found in the insoluble fractions of the cell lysates.

The purification of the *E. coli* Shuffle T7 express produced DocD-CBM1 fusions was attempted by affinity purification in an Avicel column in the presence of ammonium sulfate. High concentrations (1 M and 0.5 M) seemed to cause precipitation of the target protein and the affinity purification was attempted with 0.25 M ammonium sulfate instead, with which the proteins remained in a soluble from through the purification. From the SDS-PAGE gels and the Western blots of the Avicel purification it could be seen that most of the protein was in the flowthrough fraction (results not shown). This indicates that the CBM1s were not expressed in *E. coli* in a functional form as they were not able to bind onto cellulose.

3.1.1.2. S. cerevisiae produced proteins

In yeast the production levels of the DocD-CBM1 -constructs were much lower than in *E. coli*. The proteins were expressed extracellularly and harvested from the cultivation media. The production levels of the shorter version of the DocD-CBM1 fusion were higher than those of the longer one in yeast (results not shown). Even still, the protein amounts of the shorter version were very low, since

the proteins could not be visualized on an SDS-PAGE gel and without the Western blotting. Even with silver staining, which usually visualizes proteins in the nanogram level, the protein could not be detected even in the concentrated supernatant (results not shown). This difference between the production levels of the shorter and longer versions could probably be attributed to the signal sequences used for the expression of the proteins. The inserted gene in pSVFDP12 (short DocD-CBM1) contained a *T. reesei* xylanase 2 leader sequence followed by a kex 2 protease cutting site for extracellular expression and cleavage of the signal sequence while in pSVFDP13 (longer version of the DocD) the signal sequence was yeasts alpha-factors (Table 2, supplementary files 4 and 5). Because of the differences in the production rates of the two different constructs, the rest of the experiments with the yeast produced proteins described below were done with the shorter protein construct.

The protein band of the yeast produced proteins on the Western blot was more the like of a smear pattern, unlike with *E. coli* produced proteins, which were seen on the blot as sharp single bands (Figures 11, 12 and 10 B). This could indicate that the yeast produced protein is glycosylated and thus varies in size. The theoretical calculated sizes for the *S. cerevisiae* produced DocD-CBM1 proteins were approximately 12 kDa, which is the same as with the corresponding protein produced in *E. coli*, and does not take into account glycosylation (Table 2, supplementary files 4 and 5). The glycosylation site mutations made in the genes coding for the proteins were supposed to prevent the yeast from overglycosylating the protein, but although the putative N-glycosylation sites had been removed by mutagenesis, there were still many O-glycosylation sites available, especially in the linker region of the protein. The effects of the N-glycosylation mutations to the function of the protein were unknown and it might be that the mutations would have affected the dockerin-cohesin interaction. In other studies it has been shown that mutations in that exact specific residue have significantly reduced the affinity between the dockerin and cohesin counterparts to each other, even by as high as three orders of magnitude (Mechaly et al. 2001). The effect of the mutations for eliminating the N-glycosylation by *S. cerevisiae,* however, remained unknown as the expression levels were too low to characterize the yeast expressed DocD-CBM1 fusions.

Buffering of the culture media resulted in a different appearance of the proteins when compared to the ones produced in the unbuffered culture media (Figures 11 and 12). In the Western blot of the buffered culture there is a clearly distinguishable smaller protein band just below the 15 kDa marker that is not found in the unbuffered cultivations. The pH of the supernatant in the buffered

cultivation was from ~ 5 to 5.5 after the cultivation while the pH of the unbuffered media at corresponding stage and time was approximately 3. The smaller band could thus probably be a less glycosylated form of the protein that is produced only at a higher pH. The buffering did not, however, seemingly enhance the production levels of the proteins.

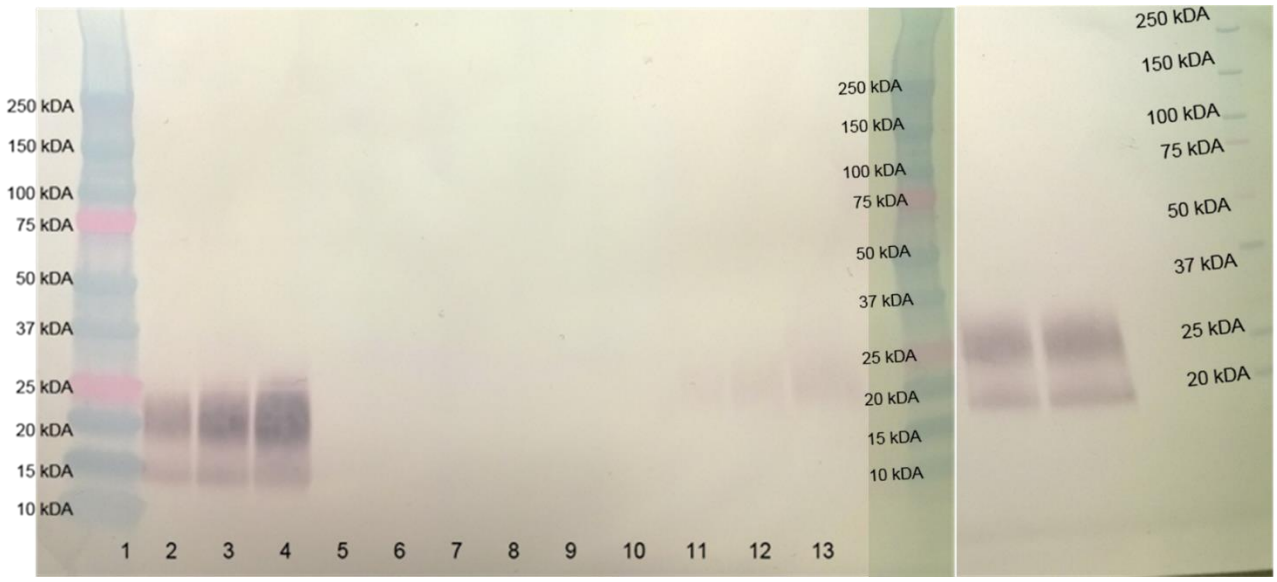


Figure 11. Buffered cultivations of DocD-CBM1 fusion construct in *S. cerevisiae*. On Lanes 2 to 4 are the 15 x concentrated yeast supernatants. On lanes 5, 6 and 7 are the 15 x concentrated samples of the "flowthrough" and two washing step of Avicel batch purification. On lanes 8 to 10 are different amounts of the 7 x concentrated 1 h elution loaded on the gel and on lanes 11 to 13 are different amounts of 15 x concentrated samples of the 1 h elution loaded on the gel. On the right the two unnumbered lanes are samples of the protein still bound to Avicel after elution with water, separated from the Avicel by heat denaturation in the presence of mercaptoethanol. The buffering of the cultivation medium produces different looking smear patterns from the unbuffered media with a clearly distinguishable smaller protein band just below the 15 kDa marker. Although some of the protein is eluted out of the Avicel most of the protein is still bound to it after elution as seen in the unnumbered lanes on the right, and so the buffering does not affect the binding or elution of the proteins out of the Avicel.

The Avicel affinity purification of the DocD-CBM1 fusions from the yeast supernatants were done in a batch mode. The purification results of the proteins from the buffered and unbuffered culture media in the Avicel batch purification can be seen in Figures 11 and 12, respectively. It seems that the binding in the presence of 1M ammonium sulfate and subsequent eluting with water worked to some extent and some protein was eluted out of the Avicel during the first hour of elution. The additional elution O/N did not enhance the elution rates much. Although a lot of protein was still bound to the Avicel after elution and was therefore lost during the purification, this also confirms that the production of a functional CBM1 that is able to bind cellulose was achieved in yeast. Much of the protein was, however, also still washed out with the flowthrough in the unbuffered cultivation

for an unknown reason, which diminishes the yields of the target protein. In the buffered cultivation it would seem that not as much of the protein was lost in the flowthrough but nevertheless still most of the proteins were bound to the Avicel after elution. The binding of the proteins to Avicel was done at 4 ˚C and eluting at room temperature to enhance the elution rates of the proteins according to Sugimoto et al. (2012). The elution of the proteins out with or in the presence of other substances like triethylamine could possibly be used to enhance the elution yields (Barak et al. 2005, Karpol et al. 2008), but this was not attempted in this thesis work.

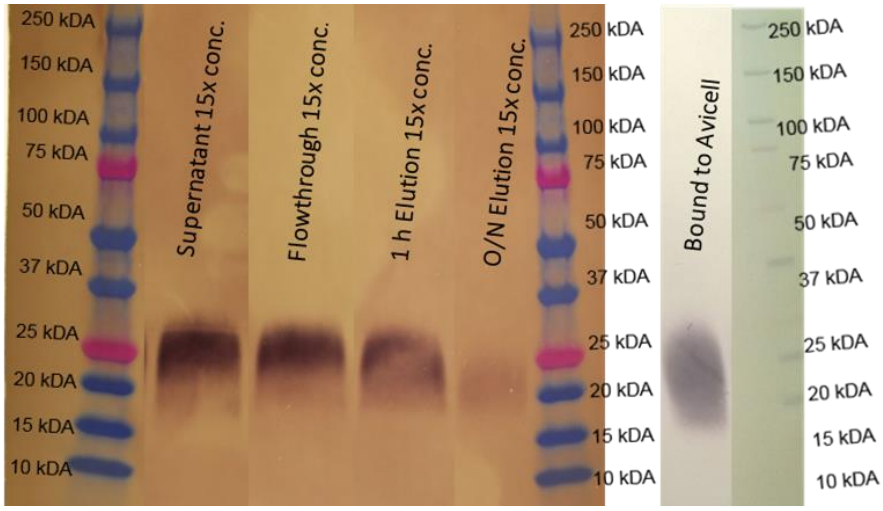


Figure 12. Unbuffered yeast cultivation supernatant batch purification. Production of the fusion proteins in an unbuffered culture media resulted in a protein smear pattern that is distinct of that produced in a buffered media (Figure 11) and no smaller, less glycosylated band can be seen in the samples taken from the unbuffered cultivation. Purification of the proteins from the unbuffered cultivation by affinity purification with Avicel in the presence of 1 M ammonium sulfate and subsequent elution with water worked to some extent and the proteins are released from the Avicel to some extent after 1 h of elution. Additional elution for O/N did not enhance the elution rate significantly. Much of the proteins were still bound to the Avicel even after the O/N elution ("Bound to Avicel" –lane), indicating the cellulose affinity purification method was not very efficient but also showing that the production of a functional CBM1 in *S. cerevisiae* was successful.

3.1.2. DocS-CBM3 fusion proteins

Expression of the Ct CelS dockerin with the Ct CipA CBM (different DocS-CBM3 fusion constructs with different linkers, Table 2) in *E. coli* was very successful. The fusion proteins containing a His-tag in either C-or N-terminus were analyzed from the cell lysates of the expression host *E. coli* BL21 (DE3) on SDS-PAGE and Western blotting using His-tag antibody (Figure 13).

DocS-CBM3 construct 2 showed in Western blot a series of smaller protein bands below what was assumed to be the intact fusion protein at ~27 kDa level (TMW ~27 kDa) indicating proteolysis (Figure 13, lanes 2 to 5), while DocS-CBM3 construct 1 showed only a single band. Since, however, multiple strong bands can be seen even for the construct 1 on the SDS-PAGE gel (Figure 13, lanes 6

to 9), when compared to the control strain (Figure 13, lanes 10 to 13), this indicates that proteolysis was also happening with the construct 1.

The reason why there are only single bands visible in the construct 1 cell lysates (Figure 13, lanes 6 to 9) in the blot is that the His-tag in this construct is located at the N-terminus of the protein (at the dockerin end), while with construct 2 and the other DocS-CBM3 -constructs, it is located at the C-terminus (at the CBM end)(Table 2). This information combined with the band sizes and the fact that the CBM3 constitutes the major part of the protein (154 amino acids, TMW ~17 kDa) compared to the dockerin (68 amino acids, TMW ~7.5 kDa) allows us to deduce that the proteolysis cleavage sites are in the dockerin or the linker part of the fusion protein, not in the CBM.

Different linker regions between the DocS and the CBM3 were tested to overcome the proteolysis issue (constructs 3-5, Table 2). A His-tag affinity purification was done for the DocS-CBM3 fusion constructs 1 to 5 with similar protocols, only with a slight modification in elution gradient length. The affinity purification results of each DocS-CBM3 fusion construct can be seen in Figure 14. 6 purification fractions for each of the proteins were chosen to be pooled together for further studies according to the gel images. With all of the constructs from 2 to 5 there can be seen multiple sized proteins that are eluted out of the column at the same time. These bands are presumably the different sized protein bands of the proteolyzed target proteins. The theoretical molecular weights (TMWs) for the DocS-CBM3 proteins were approximately 27 kDa, varying to some extent between the different versions of the proteins (Table 2, supplementary files 6 and 7). The ratios of the different sized bands compared to each other seem to vary a little between the different protein constructs, but the proteolysis in nevertheless quite a big problem with all of them. The five different constructs were made with the different linkers in order to avert the possible proteolysis problem that was thought to occur in the linker region. The hypothesis was that the fungal linker sequence originating from *T. reesei*, might be more susceptible to proteolysis in *E. coli* than the bacterial linker peptides, used in constructs from 2 to 5 (Table 2, supplementary file 7).

Since the purified protein band sizes imply that a peptide of some 3 kDa is cleaved “first” or as the smallest part followed by a second band of approximately some 6 kDa smaller than the largest intact protein (additionally followed by some other even smaller bands), this indicates that the proteolysis problem would not be in the linker region of the protein alone, but happens independently also in the dockerin part of the fusion (Figure 14). Similar kinds of problems when producing *Clostridium* spp. dockerin containing proteins in *E. coli* have been observed and reported earlier on several

occasions (Fierobe at al. 1991, Murashima et al. 2001, Carvalho et al. 2003, Caspi et al. 2006). In those studies the exact cleavage sites were not identified, but the results would suggest multiple cleavage sites in the fusion proteins and also in the dockerin domains of the fusions. Carvalho et al. (2003) reported problems in producing dockerin domains as single domains (or independent individual entities) in *E. coli* due to protein degradation in the cells and needed to produce the proteins from plasmids expressing both the dockerin and the cohesin domains in order to increase the stability of the dockerin domain to be able to purify it. These result indicate that there would be some *E. coli* protease susceptible proteolysis sites within the dockerin sequence and that these cleavage sites could act even as the initial cleavage sites from where the proteolysis might then proceed further. This may also be the case with the DocS-CBM1 fusion proteins used in this thesis, as seen as the multiple bands on the SDS-PAGE gel and the Western blot (Figure 10).

The production levels of all the DocS-CBM3 -constructs were, despite the proteolysis problem, nevertheless sufficiently high for all the proteins to work with. The production and purification rates for the different proteins seemed to vary to some extent between the constructs with the construct 2 having the biggest production rates. This could be seen even on different cultivations with varying induction times. The purification elutes contained higher concentrations of the protein with construct 2 than with other constructs according to both SDS-PAGE gel images and NanoDrop values (results not shown). With construct 1 the amount of the total protein in the elution fractions consists mostly only of the whole length, unproteolyzed or intact fusion protein, while with the other constructs also the proteolyzed proteins were present, which affected the protein concentration measurement values. This is why the production levels of the construct 1 were not directly comparable with the other constructs. With some of the protein constructs there can be seen some slight traces of some other contaminating proteins also in the elution fractions in Figure 14. When the images' coloring and resolution were changed, some contaminating trace protein could be seen in almost all of them. Despite this, most of the proteins in the elute fractions seem to consist of the DocS-CBM fusion construct and the His-tag purification seems to have worked very well.

The proteolysis of the DocS-CBM3 fusion proteins seemed to be independent of the induction and cultivation times, since even with induction times of 3 hours and cell densities (OD_{600}) of 0.12 to 0.44 when induced the proteins still suffered from proteolysis (in more or less the same proportions)(results not shown). Even when the protease inhibitor (cOmplete, EDTA free protease inhibitor (11873580001, Roche Diagnostics GmbH, Germany)) was added in 5 times excess into the cell lysis buffer mix the proteins were proteolysed. This indicates either that the proteolysis is already happening during the cell growth phase inside the cells, or that the proteases responsible for the proteolysis are not inhibited by the inhibitor we used, since e.g. metallo- and aspartic proteases are not inhibited by the cOmplete EDTA-free protease inhibitor as notified by the manufacturer of the product. Thus other protease inhibitors could perhaps be used instead of or in addition to cOmplete EDTA-free.

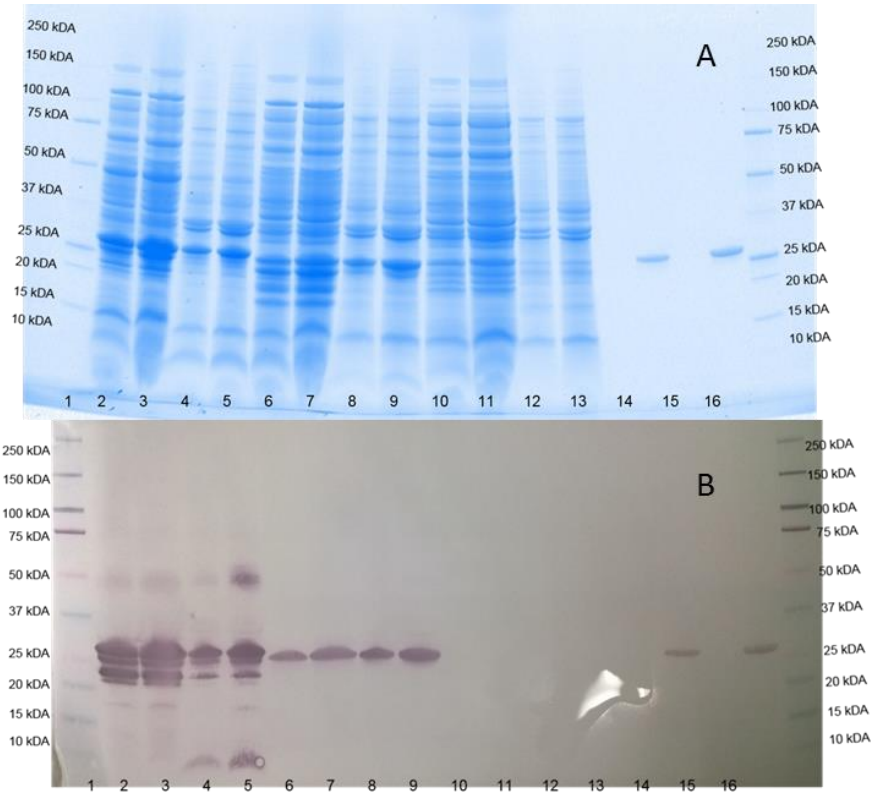


Figure 13. Production of the DocS-CBM3 fusion constructs 1 and 2 in *E. coli*. A) Cell lysates of *E. coli* BL21 (DE3) producing either the DocS-CBM3 fusion construct 2 (lanes 2 to 5) or construct 1 (lanes 6 to 9) have strong and distinguishable protein bands in the area of ~27 kDa to ~17 kDa compared to the control strain (lanes 10 to 13). The soluble fractions of the production strain cell lysates (lanes 2, 3, 6 and 7) contain a series of multiple strong bands compared to the insoluble fractions (lanes 4, 5, 8 and 9) which only contain one or two stronger bands, suggesting high proteolysis in the soluble fraction. Imaged with Criterion Stain Free Imager. B) Cell lysates visualized by Western blotting with an anti-His-tag –antibody. Multiple bands can be seen in the Docs-CBM3 -construct 2 containing *E. coli* cell lysates (lanes 2 to 5), while only single bands for the construct 1 (lanes 6 to 9) are visible. The construct 2 contained the His-tag at the C-terminus as an extension of the CBM, while the construct 1 contained the His-tag at the N-terminus, preceding the dockerin domain.

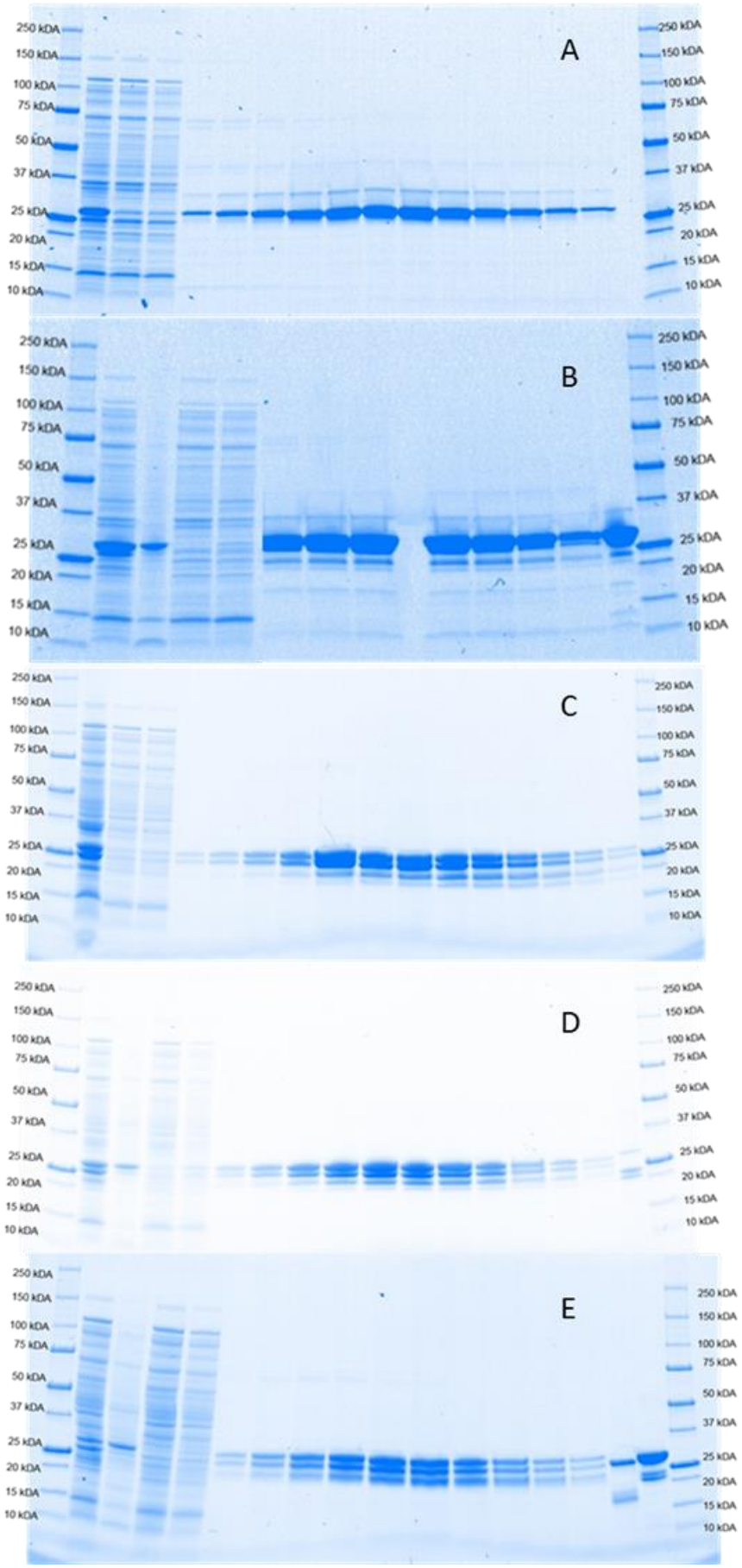


Figure 14. Purification results of DocS-CBM3 fusion constructs 1 to 5 respectively from top to bottom. In Figures B, D and E the wells consist of cell lysate soluble fraction, cell lysate insoluble fraction, purification flowthrough, purification wash, elution fractions and control protein(s), respectively from left to right. In Figures A and C the insoluble fraction of cell lysate and the control proteins are missing. In the cell lysate soluble fractions the target protein can often be seen as multiple strong bands while in the insoluble fraction the protein is usually seen as only one band. The elution fractions usually contain a lot of protein in multiple purified bands, which infers proteolysis. With construct one (Figure A), however, there can be seen only one clear strong band in the elution fractions, which is supposed to be the intact whole DocS-CBM3 fusion protein. In the construct 1 the His-tag is located at the N-terminus, while in the other constructs the His-tag is in the C-terminus of the protein. Because of this, the whole spectrum of proteolyzed proteins are purified by the affinity purification with the constructs from 2 to 5. With some of the protein elution fractions there can also be seen some slight traces of some other contaminating proteins in the mixture. Figures imaged with Criterion stain free imager.

3.1.3. Purity of the proteins

The purity of the DocS-CBM3 fusion constructs used in this study can be seen in Figures 14 and 16. The purity of the TeCel7A- cohesin fusion and all the control proteins used in this study can be seen in Figure 15. The TeCel7A-cohesin fusion (lane 3, Figure 15) contains quite much of contaminating proteins which can be seen as the smaller bands at the level or below the 50 kDa protein marker on the gel. The *S. cerevisiae* produced TeCel7A core enzyme, used as a control enzyme for the TeCel7A-cohesin fusion on lane 4 is quite glycosylated, which can be seen as the smear pattern on the lane over the more distinct protein band. The different *S. cerevisiae* produced peptide linked TeCel7A-CBM3 fusions (lanes 5, 6 and 7) contained various levels of glycosylation and contaminations. Of these the protein preparation on lane 6 seems to be the purest and was therefore chosen to be used in the further experiments as a control for the enzyme complex and is later on referred in this thesis as the *S. cerevisiae* produced TeCel7A-CBM3 fusion.

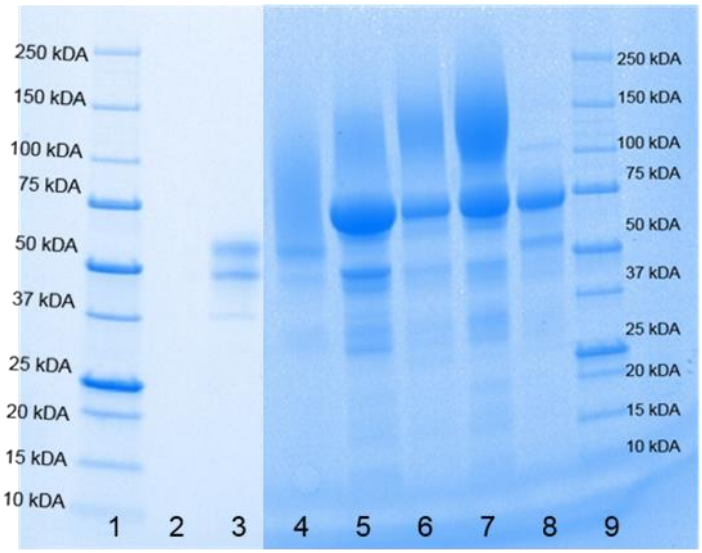


Figure 15. SDS-PAGE image of the control enzymes and the cohesin-catalytic domain fusion protein. Lane 3: TeCel7A-cohesin fusion; lane 4: *S. cerevisiae* produced core of TeCel7A; lanes 5, 6 and 7: differentially glycosylated *S. cerevisiae* produced TeCel7A-CBM3 fusions (lane 6 used in further studies).

3.2. Characterization of the proteins

3.2.1. Storage stability of the fusion proteins

The storage stability studies revealed that the proteolysis of the His-tag purified DocS-CBM3 fusion proteins still proceeded during the storage of the proteins at 4 ˚C (Figure 16). The smaller protein bands appearing in the purified proteins seem to degrade further, and at a faster rate than the intact whole proteins as can be seen in Figure 16 as the smaller protein bands weakening and even smaller bands emerging still in the samples. The intact or whole fusion proteins, seen as the biggest protein

band seem to nevertheless last quite well even at 4 ˚C temperature, which is good considering their storage, use in further studies and in applications performed even for longer time periods. Furthermore, the proteins stored at -20 ˚C seem to remain at the same condition as they were when they were purified when compared to the purification elution fraction images (Figure 16 B). This is crucial considering the further studies, because it allows for the acquisition of constant and reproducible results with the same protein preparations performed at different times. Storage of the proteins for longer periods of time than 22 days was, however, not studied. Storage at -80˚C would also be a viable option if a more certain or durable storage for much longer periods of time should be needed. Since the His-tag -purified DocS-CBM3 fusion construct 1 seemed to contain mostly the whole, non-proteolysed fusion protein (Figure 14) and it maintained in that form during the storage at -20 ˚C and since the DocS-CBM3 construct were all nearly the same, the activity studies and rest of the characterization of the proteins was decided to be done with this construct.

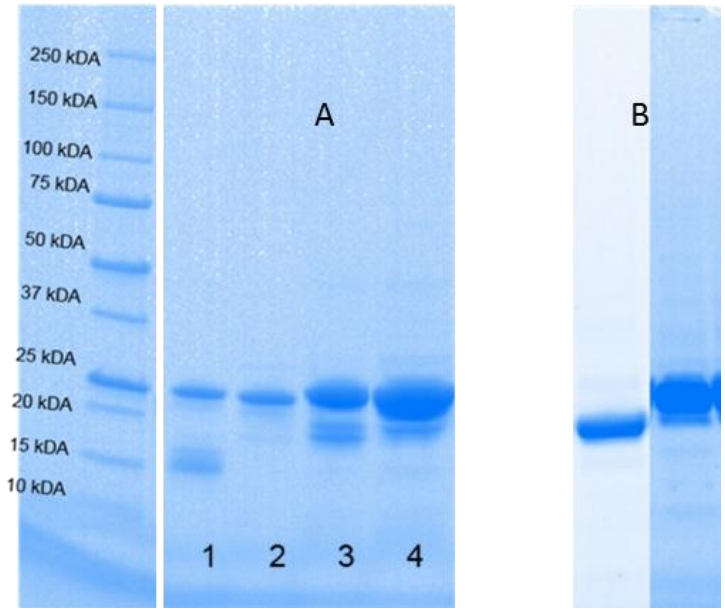


Figure 16. Storage stability of the DocS-CBM3 fusion construct at different temperatures. A) On lanes 1 and 2 are the samples of construct 1 and on the lanes 3 and 4 the samples of construct 2 stored at 4 ˚C (lanes 1 and 3) or at -20 ˚C (2, 4) for 19 days. In the samples stored at 4 ˚C, the proteolysis has proceeded further which can be seen as the smaller protein bands weakening and even smaller bands emerging. At -20 ˚C the proteolysis does not seem to proceed further when compared to the images of the same proteins in the purification elution fractions (Figure B). B) His-tag-purification elution fractions of constructs 1 and 2 respectively. Imaged with Criterion stain free imager.

3.2.2. Complex formation

Complex formation between the cohesin and dockerin fusion proteins with varying incubation times was visualized on non-reducing SDS-PAGE without mercaptoethanol in the loading buffer and without heat denaturation of the proteins. The complex formation could be already seen with incubations times of approximately 10 minutes in RT as seen in Figure 17. The complex is formed

and can be seen as a distinct band above the two bands found in the TeCel7A-cohesin fusion purification preparation and in the mixture of the two protein counterparts (Figure 17, lanes 5 and 7). The size of the complex on the non-reducing SDS-PAGE would seem to be approximately 70 kDa according to the gel image, although in non-reducing SDS-PAGE the proteins do not necessarily run directly according to their molecular weights. The complex formation between the two counterparts does not seem to be 100% as there can be seen a lot of the both proteins separately in the gel. This however was thought to result from the gel driving conditions. The stability of the complex may have been affected by several features such as the elevated temperature of the gel during the gel drive (caused by the high voltage), the application of the native loading buffer containing glycerol and the SDS in the gel drive buffer (which could have interfered with the binding and hydrogen bonding of the dockerin and cohesin) (Lamed et al. 1983). Although the optimal ratios for the different protein counterparts and the saturation point for total complex formation of the TeCel7A-cohesin counterpart could not be seen in the gels, the complex formation after 10 minutes at RT was nevertheless seen and proved and the studies were decided to be taken onward into the activity studies of the proteins. The complex formation results are well in line with former studies (Lytle et al. 1996, Mechaly et al. 2000) in which the complex formation between cohesin and dockerin has been shown in 10 minutes or less. Lytle et al. (1996) also reported that complex formation did not proceed further after 10 minutes of incubation.

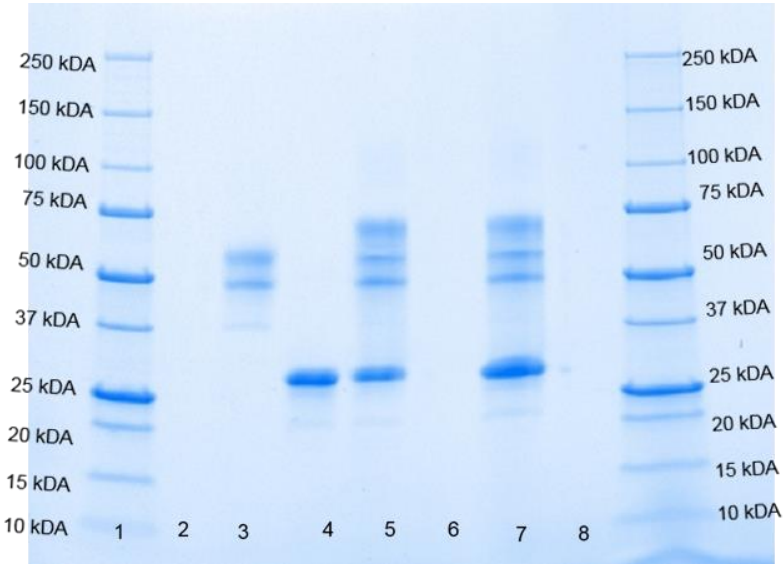


Figure 17. TeCel7A-cohesin/DocS-CBM3 -complex formation. An image of the two protein counterparts incubated with each other in RT for 10 minutes. On lane 3: the TeCel7A-cohesin fusion, on lane 4: DocS-CBM3 fusion, on lanes 5 and 7 mixtures of the two protein preparates and complex formation between them. The complex can be seen in the mixture lanes above the TeCel7A-cohesin -bands, just beneath the 75 kDa marker band. 4x Native loading buffer stock was added in 1:5 volume in the protein dilutions. Running of SDS-PAGE with ≤ 200 V. Imaging of the gel with Criterion stain free imager.

3.2.3. Soluble substrate hydrolysis

The activity of the catalytic domain-cohesin fusion (TeCel7A-cohesin) was measured on soluble substrate (MULac) to verify that the cohesin domain does not reduce the catalytic efficiency of the Cel7A domain. The MULac activity of the TeCel7A-cohesin was compared to the activity of *S. cerevisiae* produced control proteins TeCel7A (core without CBM) and TeCel7A-CBM3 (Figure 18). The activities for the different enzyme preparations are for 0.2 µM total protein concentration samples, which contained varying amounts of differentially glycosylated or contaminating proteins, as seen from Figure 15. The specific activities for the different protein preparations can be found in Table 3. The activities of the protein preparations are within 11% range of the control protein preparation of Sc produced TeCel7A-CBM3. Because the activities were in this quite near proximity to each other the loading of the proteins for the Avicel hydrolysis was decided to be done according to the molar concentrations of the protein in the protein preps, acquired by their absorbances at 280 nm as with the soluble substrate hydrolysis protein preps. Since the TeCel7A-cohesin fusion enzyme preparation was nearly around 10% less active than the TeCel7A-CBM3 fusion used as a control enzyme for the complex in Avicel hydrolysis and furthermore around 20% less active than the core of TeCel7A used as the reference for the TeCel7A-cohesin fusion, this needs to be taken into account when assessing the results obtained from the Avicel hydrolysis.

The MULac specific activity for the TeCel7A core was similar to that reported by Voutilainen et al. (2014). The specific activity for the TeCel7A-CBM3 fusion reported here, however, differed from the values reported in the other study, being much lower than those reported earlier. The values for the TeCel7A-CBM3 fusion reported here are lower than those of the TeCel7A core while with Voutilainen et al. (2014) they were the other way around. This difference may result from the use of different batches or purification fractions of the enzymes or even from the storing of the enzymes for long periods of time. Nevertheless the enzymes specific activities were in the same order of magnitude with the core enzyme activities exhibiting similar values with each other confirming the enzymes were still functional.

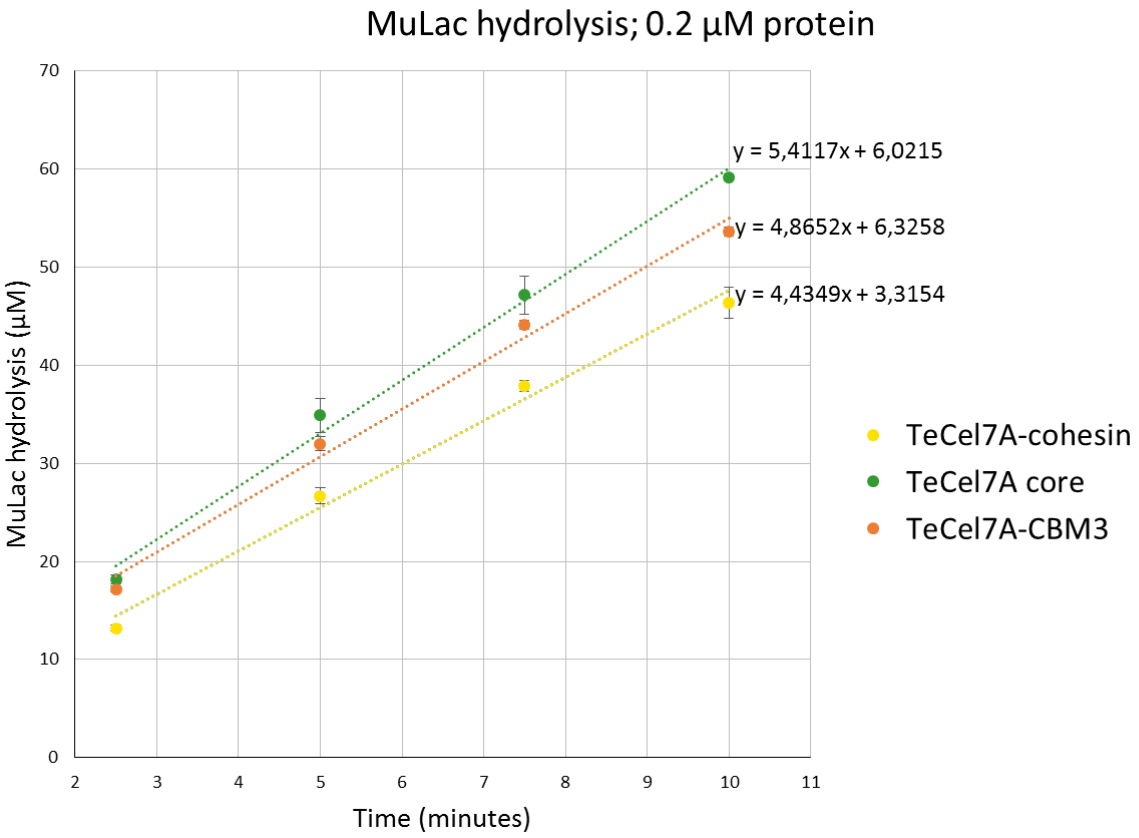


Figure 18. MULac hydrolysis values for each of the enzyme preparations of catalytic domains used in this study. The hydrolysis values of MULac depicted here are mean ± S.D. of triplicate measurements for 0.2 µM protein preparations for different time periods compared against a MU standard.

Table 3. Specific activities of the protein preparations. The specific MULac hydrolysing activities of the protein preparations against a MU standard. The specific activities are depicted here as the amount of MULac molecules one enzyme is able to hydrolyze in one minute. The Sc produced TeCel7A-CBM3 has been chosen as the relative activity reference enzyme to which the other ones are compared to.

Protein preparation	Specific activities	Relative activity (%)
TeCel7A-cohesin (Sc produced)	22.2 min^-1	91
TeCel7A core (Sc produced)	27.0 min^-1	111
TeCel7A-CBM3 (Sc produced)	24.3 min^-1	100

3.2.4. Thermostability

3.2.4.1. DocS-CBM3 fusion

The circular dichroism (CD) temperature ramping spectra of DocS-CBM3 fusion in 10 mM Tris-HCl buffer (pH 7) supplemented with 5 mM $CaCl_2$ can be seen in Figure 19. The protein has a typical β-

sheet secondary structure spectrum. Changes in the spectra of DocS-CBM3 fusion start to occur at around or after 55 ˚C and the spectrum has clearly differentiated from the spectra at lower temperatures already at 59 ˚C (Figure 19, green line). This differentiation of the spectra curves was interpreted as the unfolding of the proteins, although it does not resemble the traditional unfolded spectra curve of proteins. When heated to higher temperatures (Figure 20, red and brown lines and Figure 19, red line) the spectra differentiated from the RT spectra (Figure 19, black line) even more. The original RT spectra of the protein, however returned when the sample was cooled back down to RT (Figure 19, green line), even after a few minutes retention time at temperatures from 85 to even 90 ˚C. Returning of the original spectra was interpreted as refolding of the proteins back to their original fold. This assumption was also supported by the fact that there wasn't any precipitation of the proteins found in the cuvettes during or after the heating. This same kind of behavior of the protein was also observed in 10 mM NaAc buffer (pH 5), supplemented with 5 mM $CaCl_2$.

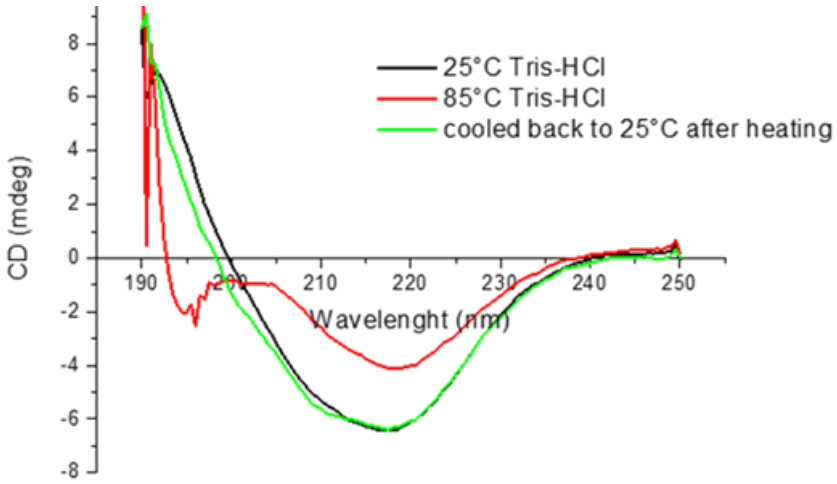


Figure 19. DocS-CBM3 fusion circular dichroism spectra. The spectra of the protein is shown in RT before heating (black line), when heated to 85 ˚C (red line) and when cooled back to RT again (green line). The spectra in the RT are similar to each other while the heated or denatured protein sample has a very different looking spectrum. A 3 µM protein concentration in Tris-HCl buffer (pH 7) supplemented with 5 mM calcium was used in the measurements.

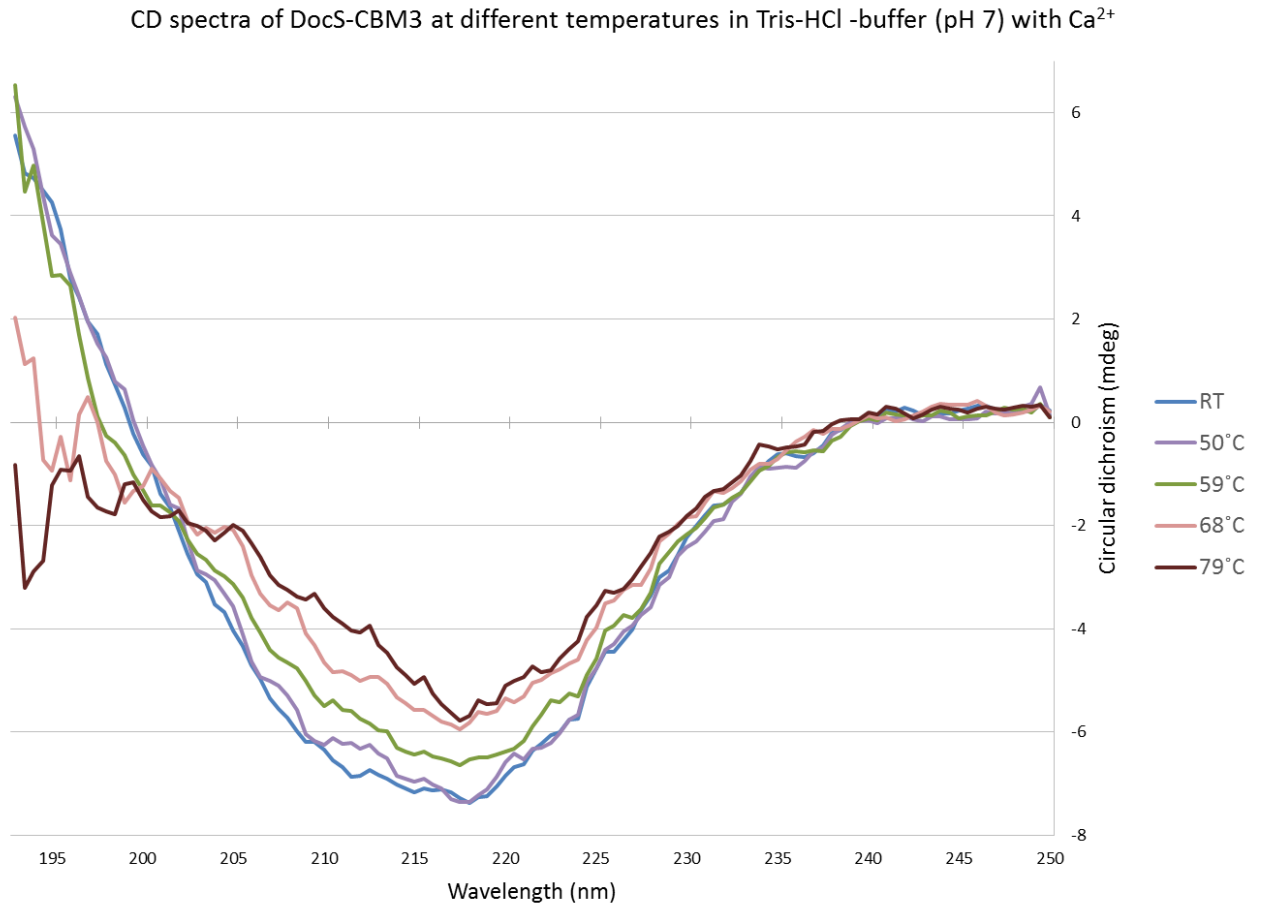


Figure 20. The CD spectra of the 3 µM DocS-CBM3 fusion construct 1 at different temperatures in Tris-HCl buffer (pH 7) supplemented with 5 mM calcium. The spectra of the proteins remain nearly unchanged at low temperatures (blue and purple lines) until temperatures of 55 ˚C to 60 ˚C (green) at which temperature it gradually starts to change and differentiate from the lower temperature spectra. The spectra keeps changing all the way up to some 80 ˚C by rising in the 200 to 235 nm wavelengths and by lowering in wavelengths lower than that (red and brown lines).

CD spectrum of DocS-CBM3 was also recorded without added calcium in the buffer. The shape of the spectrum in room temperature was slightly different (Figure 21, blue line) as compared to the spectrum in the presence of added calcium (Figures 19 and 20). Temperature induced changes in the spectra could be seen starting at around 70 ˚C. Also some precipitation of the protein could be seen in the cuvette after the temperature ramping, and the original RT spectrum of the protein did not return after cooling the sample back down to RT (data not shown). Similar temperature induced behavior of the spectra could be seen with the proteins in the buffer without added calcium and with ~300 µM EDTA in the solution (data not shown). The results imply that the fusion protein loses at least some of its calcium atoms as it unfolds at high temperatures, precipitates and cannot thus

refold back to its original form in low calcium concentrations in the buffer, i.e. without the added extra calcium.

These results are in line with former studies of the dockerin unfolding and refolding and binding activity in the presence of EDTA (Stahl et al. 2012). In those studies it was discovered that when subjected to mechanical pulling force in the presence of EDTA the dockerin loses its ability to refold and thus bind its cohesin counterpart again in consecutive pull and ease series. When returned to a buffer containing calcium and no EDTA, however, the protein recovers its ability to bind the cohesin again. This means that the folding of the dockerin is reversible as long as there is Ca^{2+} available in the solution (Stahl et al. 2012). The reason why the precipitation of the proteins happens only in the buffer where no added calcium is present remains unanswered. It may be that only when there is no added calcium in the buffer the calcium is able to dissociate from the protein thoroughly resulting in precipitation of the proteins, while with excess calcium present the dissociation is not complete.

Since both the dockerin and the CBM3 domains contain calcium, it is very hard to say which of the domains is responsible for the different looking CD-spectra in the presence and absence of the calcium. In order to determine the reasons for these calcium dependent differences and the functionality of the different domains at different temperatures, some additional tests, such as cellulose binding or complex formation tests would need to be done with the heat treated proteins. Since the proteins nevertheless seem to refold back in the calcium-supplemented buffer even after heating to high temperatures it may be advantageous to always have an excess of calcium in the buffers when using this protein at high temperatures.

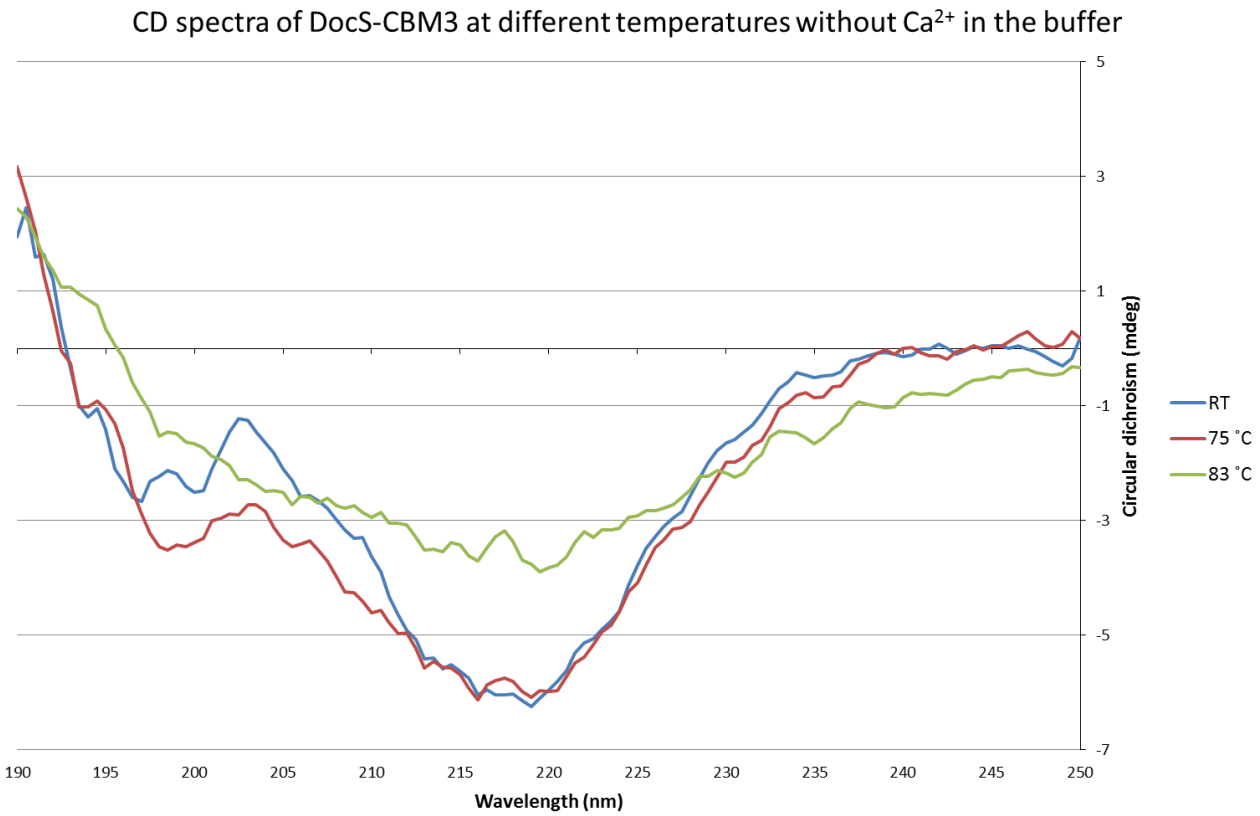


Figure 21. The CD spectra of the 3 µM DocS-CBM3 fusion protein construct 1 in a temperature ramp from RT to 83˚C without added calcium in the buffer. The protein spectra began to change at around temperatures of 70 ˚C and higher and the differentiation in the spectra can be clearly seen at 75 ˚C (reddish brown line). At even higher temperatures the spectrum has completely changed compared to the lower temperatures (green line). An additional decline in the spectra can be seen in the low temperature spectra around the area of 195 nm to 203 nm compared to the RT spectra with calcium in the buffer (Figures 19 and 20). The decline in the spectra cannot be seen anymore at high temperatures, compared to the spectra with calcium in the buffer.

3.2.4.2. TeCel7A-cohesin fusion

The CD spectra of the TeCel7A-cohesin fusion in 10 mM NaAc buffer supplemented with 5 mM $CaCl_2$ is also a typical β-sheet spectrum at low temperatures (Figure 22). The spectra of the protein started changing at around 67 ˚C (Figure 22, reddish brown), which implies protein unfolding. At temperatures over 76 ˚C the protein precipitated and the precipitated, which can be seen as the loss of spectra (Figure 22, turquoise line) and as cloudy white aggregate in the sample cuvette.

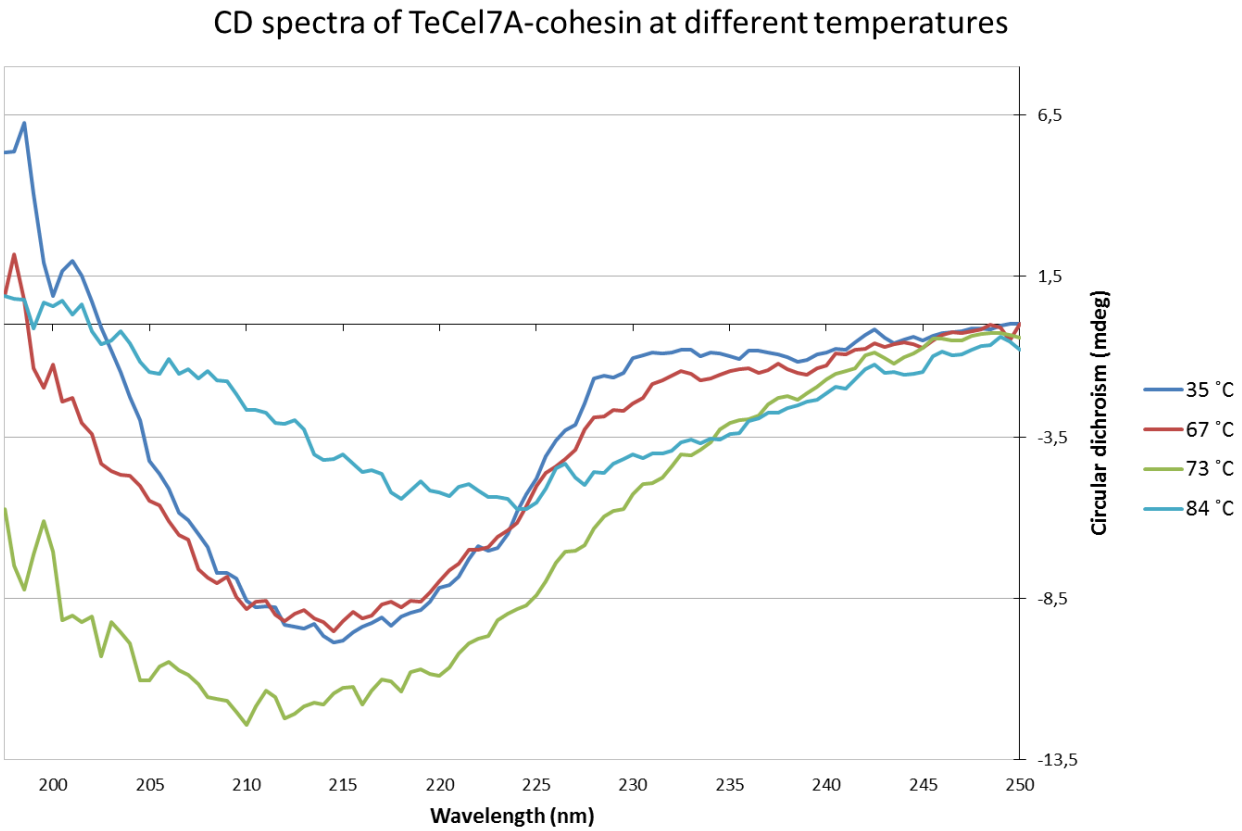


Figure 22. The CD spectra of the TeCel7A-cohesin fusion at different temperatures. The spectra of the protein at lower temperatures (purple) resembles a typical β-sheet protein spectrum. The spectra begin to change at temperatures of ~67 ˚C (reddish brown) and the spectra continues to drop (green) until up to 76 ˚C (not shown), which can be interpreted as protein unfolding. After 76 ˚C the protein started precipitating, resulting in the loss of spectra at higher temperatures (turquoise). CD spectra for the proteins were measured in 10 mM NaAc -buffer (pH 5) supplemented with 5 mM $CaCl_2$ and with 3 µM protein concentration.

3.2.4.3. TeCel7A-cohesin/DocS-CBM3 -complex

The temperature behavior of the complex (TeCel7A-cohesin/DocS-CBM3 (construct 1)) was similar to the TeCel7A-cohesin fusion. The spectra of the complex started changing at around 67 ˚C (red line, Figure 23), which is the same as with the TeCel7A-cohesin fusion. The protein also precipitated at higher temperatures, resulting in the loss of spectra (green line, Figure 23).

Since the spectra of the complex, however, remained unchanged until 67 ˚C, this could imply that the complex formation stabilized the DocS-CBM3 fusion part of the complex, which alone unfolded already at around 60 ˚C. From these results it also might be said that it could be possible that the complex would still be able to function even at 70 ˚C in the hydrolysis of cellulose since the substrate may stabilize the protein complex and enable hydrolysis even at higher temperatures.

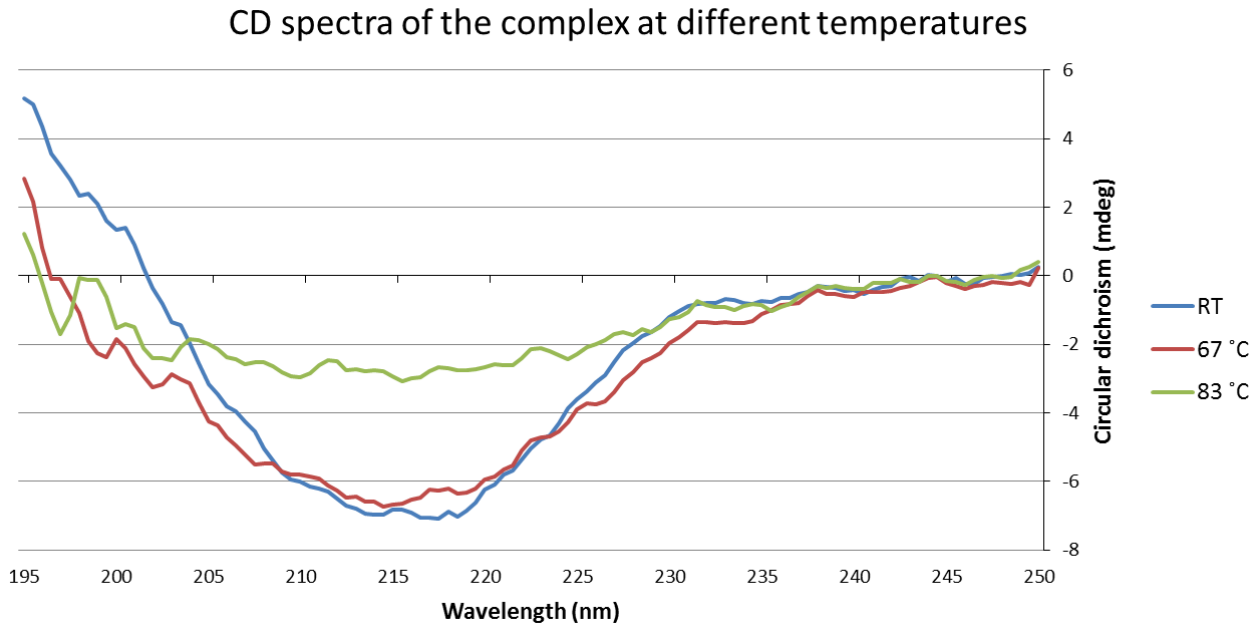


Figure 23. The CD spectra of the protein complex at different temperatures. The lower temperature spectra of the protein complex (blue line) started changing at around 67 ˚C (reddish brown line). At high temperatures the protein precipitated, which also resulted in the loss of the CD spectra (green line). The spectra and the temperature behavior of the TeCel7A-cohesin/DocS-CBM3 –complex resemble much those of the TeCel7A-cohesin fusions. CD spectra for the proteins were measure in 10 mM NaAc -buffer (pH 5) supplemented with 5 mM $CaCl_2$ and with 3 µM protein concentration for each of the complex counterparts.

3.2.5. Cellulose hydrolysis

The functionality and activity of the TeCel7A-cohesin/DocS-CBM3 -complex was studied by its crystalline cellulose hydrolysis capacity at different temperature. The hydrolysis results of 1% Avicel by the complex and the control proteins at different temperatures can be seen in Figures from 24 to 29. The hydrolysis rates of the complex are very similar to those of the TeCel7A-CBM3 fusion control enzyme at 50 ˚C and 60 ˚C. The hydrolysis rates of the TeCel7A-cohesin fusion alone are much lower than the complexes and similar to the levels of the catalytic domain of TeCel7A alone (TeCel7A core). This implies that the complex formation is effective and the enzyme complex is working well (without restrictions caused by the “cohesin-dockerin linker”) and the CBM3 attached to the catalytic domain via the cohesin-dockerin -“linker” enhances the capability of the enzyme to hydrolyze cellulose. The DocS-CBM3 fusions effect on the cellulose and its hydrolysis alone was also studied, but the fusion did not have any effect on the formation of cellobiose from cellulose on its own (results not shown). Although there are some differences in the hydrolysis rate of the complex and the control at these temperatures, these differences fall almost even within the error rates of the triplicate measurements. Since there also were some slight differences in the specific activities of the protein preparations as determined by their soluble substrate hydrolysis rates, this also

affects the accuracy of the results obtained from the Avicel hydrolysis. The general trend seen here however is that, at 50 ˚C and 60 ˚C the complex behaves and performs similarly to the control enzymes in crystalline cellulose hydrolysis. The hydrolysis rates of the complex and the TeCel7A-CBM3 fusion control at 50 ˚C equal to approximately 5% of total hydrolysis of the substrate and the rate increases to c. 7% with both the complex and the control when the temperature is raised to 60 ˚C.

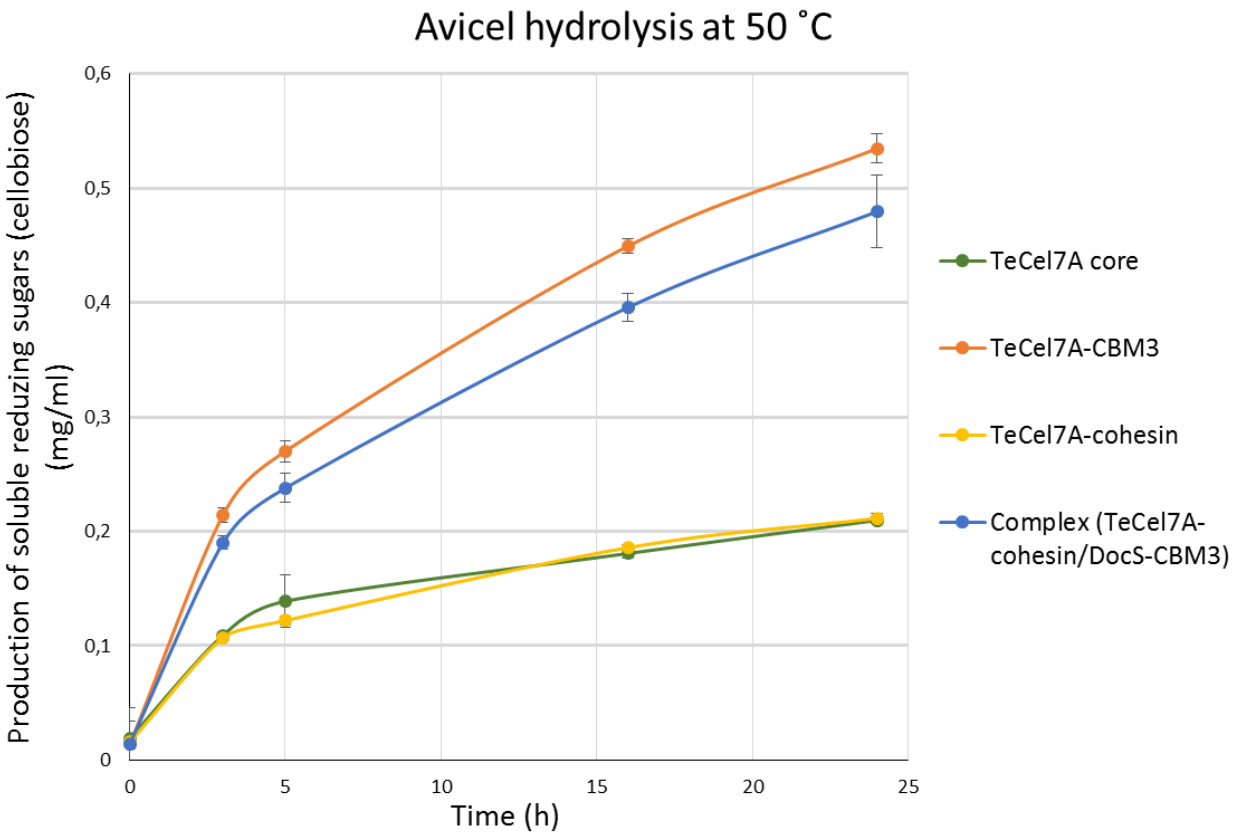


Figure 24. Avicel (1%) hydrolysis at 50 ˚C. The hydrolysis rate of the TeCel7A-cohesin fusion (yellow) is enhanced almost to the level of the TeCel7A-CBM3 fusion control protein (orange) when the DocS-CBM3 is added to the mixture and a complex is formed (blue). The hydrolysis rate of TeCel7A-cohesin fusion alone is at the level of the core of TeCel7A (green). The amount of ~0.5 mg/ml cellobiose produced during the 24 hour hydrolysis by the enzyme complex equals to ~5% of total hydrolysis of the substrate. Mean ± S.D. of triplicate measurements are shown. Enzyme concentrations in the reaction were ~0.56 µM. The accuracy of actual temperature in the samples was +-2.1 ˚C.

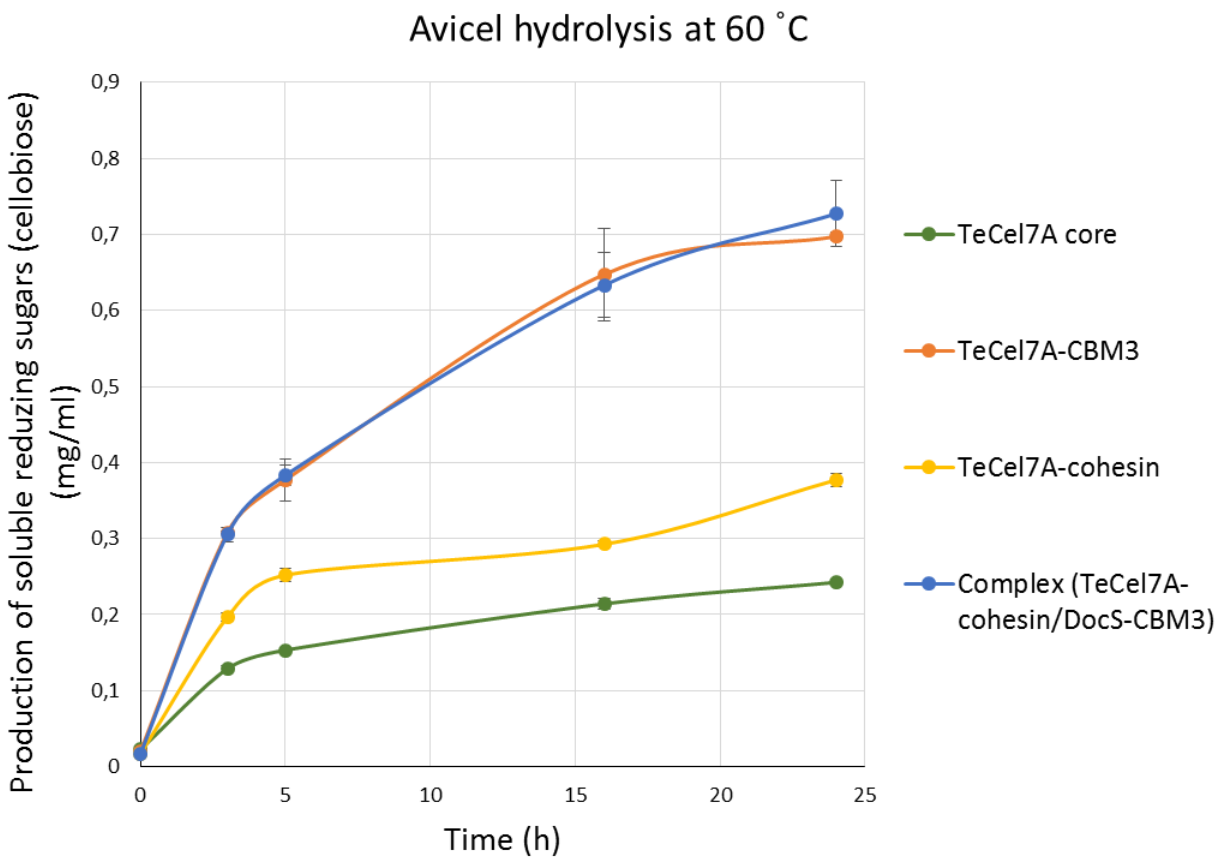


Figure 25. Avicel (1%) hydrolysis at 60 ˚C. The hydrolysis rate of the complex (blue) is almost equal to the TeCel7A-CBM3 fusion control proteins (orange). The hydrolysis rates of the TeCel7A-cohesin fusion (yellow) and the core of TeCel7A (green) are somewhat differentiated at this temperature although they both still remain much lower than the hydrolysis rate of the control fusion proteins and the complex. The amount of ~0.7 mg/ml cellobiose produced during the 24 hour hydrolysis by the complex equals to ~7% of total hydrolysis of the substrate. Mean ± S.D. of triplicate measurements are shown. Enzyme concentrations in the reaction were ~0.56 µM. The accuracy of the actual temperature in the reactions was not measured at this temperature.

At 70 ˚C, however, the hydrolysis rate of the complex drops drastically to approximately one fourth of the rate at 60 ˚C, when the drop for the control enzymes is approximately half (Figures 27, 28 and 29). The complex seems to perform better at 70 ˚C than the TeCel7A-cohesin in almost equal ratios (~2.3 times better) as in lower temperatures, which would imply that the complex is almost equally thermostable with the TeCel7A-cohesin fusion. Because the complex still performs much weaker than the TeCel7A-CBM3 fusion control, it can be deduced that the unfolding or the breaking of the enzyme complex happens after a while at these temperatures. Initially at this temperature, however, the CBM3 attaches the complex to the cellulose and allows the catalytic domain to hydrolyse the cellulose more efficiently to some extent before the complex breaks. The hydrolysis rate of the TeCel7A-cohesin fusion is thus enhanced by the addition of the DocS-CBM3 fusion to the

mix also at these temperatures. It might be that the complex formation and/or binding to the substrate may have stabilized the proteins and the protein complex to some extent. Since the CBMs have been shown to increase the amount of the catalytic domains in the substrate surface it may be that the DocS-CBM3 fusion to has conferred some stability to the complex also due to the increase of the substrate stabilizing effect for the catalytic domain conferred by the close proximity of the catalytic domain to cellulose.

The actual temperatures of the reaction mixes in the hydrolysis were lower than the set temperature of the incubator and e.g. varied between ~47.9 and 48.8 ˚C when the incubator was set to 50 ˚C temperature. The differences between the actual reaction mix temperature and the set incubator temperature at 60˚C or 70 ˚C were not measured, but the differences between the actual and set temperatures at these temperatures have probably been somewhat greater than at 50 ˚C, which need to be kept in mind when interpreting the results. The different incubations with the enzymes were performed on a few several occasions and with different incubators to verify the results.

Because the protein preparations used at the cellulose hydrolysis experiments contained some impurities and/or were differentially glycosylated the molar amounts of the intact proteins and fully active enzymes in the protein preparations may have been somewhat lower than the measured protein concentration. Since the specific catalytic activities of the enzyme preparations were nevertheless quite close to each other (at maximum ~20% difference from each other as shown by the soluble substrate hydrolysis, Figure 18 and Table 3), the proteins were loaded to the final hydrolysis mix based on the molar concentrations of the enzyme preparations and not the activity. This (20%) difference should still be taken into account when interpreting the results. Another uncertainty to the hydrolysis was brought by the complex formation. Because of the impurities in the protein preparations, the molar ratios for each of the proteins in the complex formation were not clear. The complex formation studies, analyzed by SDS-PAGE, also gave some indications that the complex formation might not be quantitative, i.e. 100%. This may have, however, been due to the SDS-PAGE running conditions. Nevertheless, to circumvent these possible problems in the complex formation, double molar concentration of the DocS-CBM3 preparation was used in the complex formation as compared to the concentration of the TeCel7A-cohesin preparation. The complex formation could also be tested in future with some other method e.g. with liquid chromatography.

The observed temperature behavior of the complex at 70 ˚C is in line with the results observed in the thermostability studies done with CD, as in those studies the complex began to unfold at 67 ˚C. The hydrolysis results also correlate well with former studies of Moraïs et al. (2016). In those studies the thermostability of the *C. thermocellum* Cel48S in complex with its corresponding cohesin in a designer scaffoldin was studied and the complex was able to withstand temperatures of 65 ˚C for 4 hours and 70 ˚C for 3 hours. In the Avicel hydrolysis results of this thesis work it can be seen that the hydrolysis does not proceed further after 3 hours at 70 ˚C, although the actual denaturing point may have been far before that. The complex has nevertheless endured the high temperatures for at least some time and enabled an enhanced hydrolysis rate for the catalytic domain for that period of time.

The hydrolysis temperature behavior of the control enzymes observed here is similar to the results reported earlier by Voutilainen et al. (2014), although in their results the drop of activity between 60 ˚C and 70 ˚C with the core and TeCel7A-CBM3 was a little bit smaller. The core of TeCel7A seems to perform better at 70 ˚C than the TeCel7A-cohesin fusion, but since the fluctuations between these proteins at temperatures of 50 ˚C and 60 ˚C were so big, the real difference in the performance at different temperatures between these two is not crystal clear. Also the specific activities between these two protein preparations differed the most (by ~20%), which may also affect the results. Nevertheless, despite of these small inaccuracies, the main observations are, however, consistent and well in line with each other and gave similar results on several consecutive performances. This indicates soundness of the results. Furthermore, because the thermostabilities for the proteins were also measured with CD, and the hydrolysis results are in line with the CD results and also the former values presented in literature (Voutilainen et al. 2014), the cellulose hydrolysis values obtained in this thesis work can be considered as trustworthy.

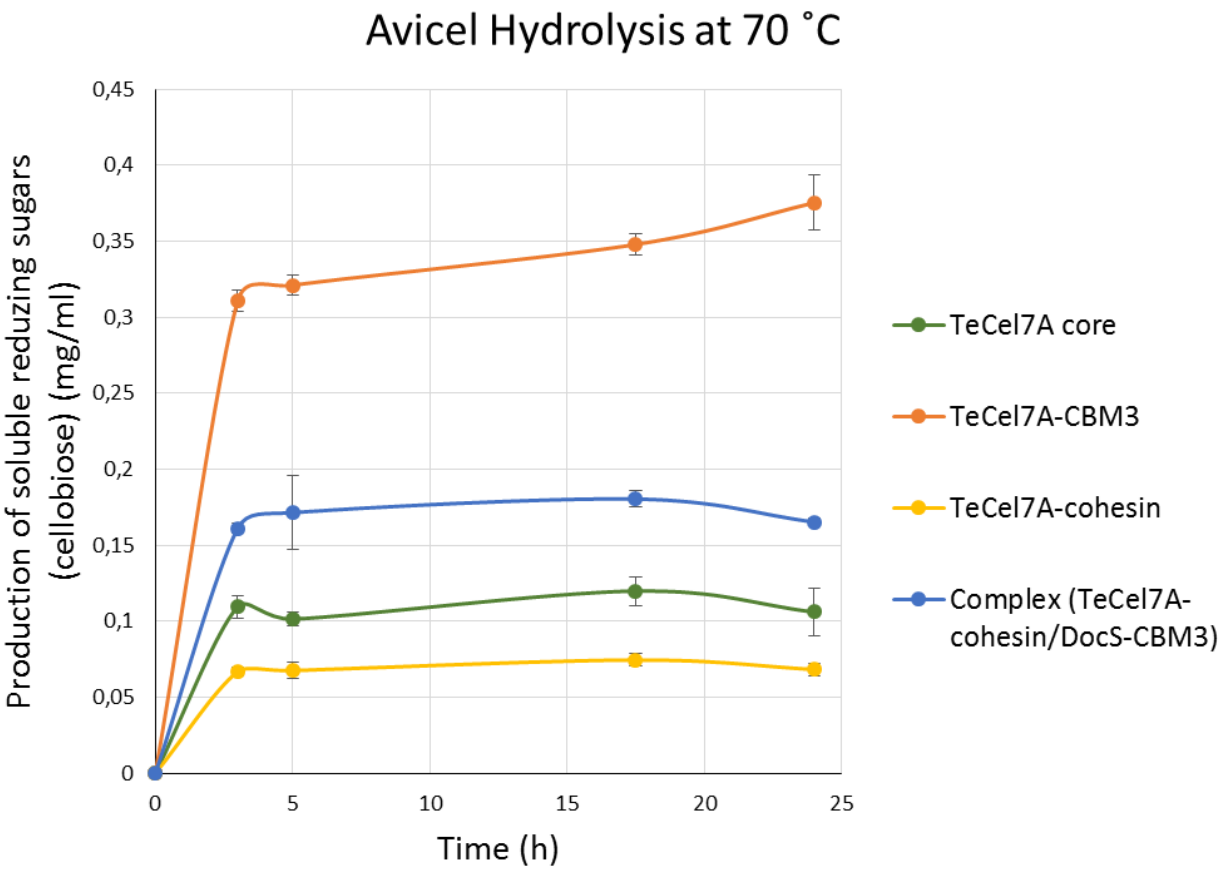


Figure 26. Avicel (1%) hydrolysis at 70 ˚C. The hydrolysis rate of the complex (blue) is much lower than that of the TeCel7A-CBM3 fusion control protein (orange), but yet still better than the hydrolysis rate of just the TeCel7A-cohesin fusion (yellow), which is the lowest of the series. The core of TeCel7A (green) performed moderately compared to the complex and TeCel7A-cohesin fusion. The hydrolysis rates compared to the ones at 60 ˚C (Figure 25 and 27) have dropped drastically since the amount of ~0.35 mg/ml cellobiose produced during the 24 hour hydrolysis by the TeCel7A-CBM3 fusion equals to ~3.5% of total hydrolysis of the substrate and the ~0.17 mg/ml produced by the complex equals to ~1.7% of total hydrolysis compared to the 7% of total hydrolysis rates at 60 ˚C. With all the proteins the hydrolysis does not proceed further much or at all after 3 hours of hydrolysis, which indicates protein denaturation due to high temperatures. Enzyme concentrations in the reaction were ~0.56 µM. Mean ± S.D. of triplicate measurements are shown. The accuracy of the actual temperature in the reactions samples was not measured at this temperature.

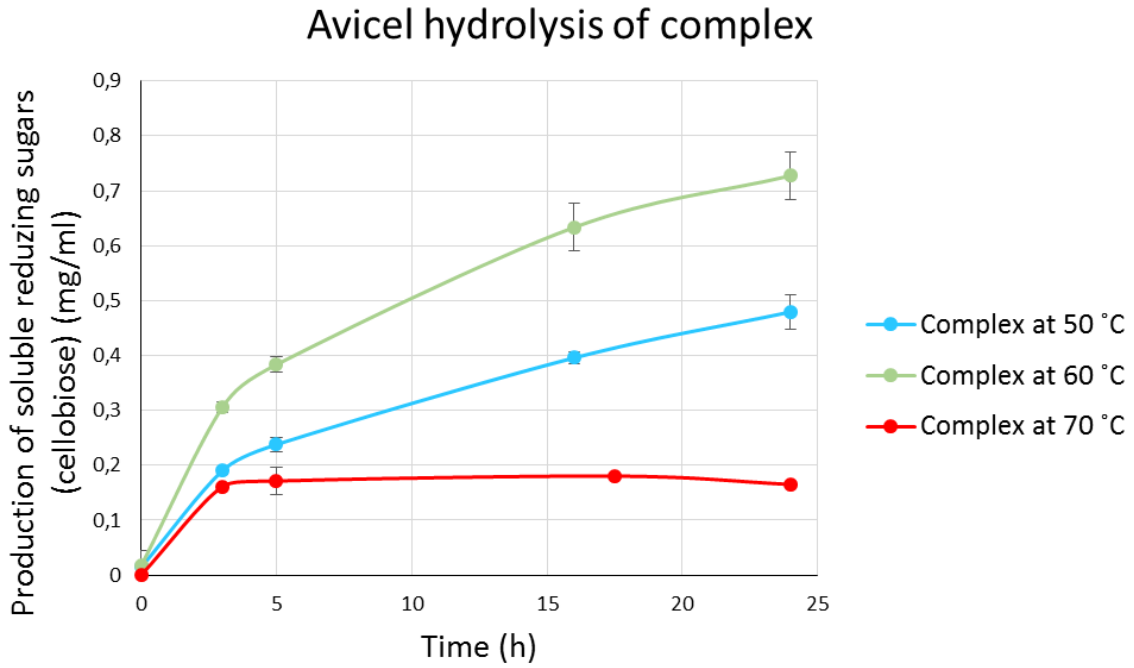


Figure 27. The Avicel (1%) hydrolysis ratios of the complex at different temperatures. The hydrolysis rate of the complex in enhanced by the temperature rise from 50 ˚C (blue) to 60 ˚C (green), by almost 50% but is then dropped to almost one fourth of the rate at 60 ˚C, when hydrolysis is done at 70 ˚C (red). The hydrolysis at 70 ˚C does not proceed after 3 hour hydrolysis, indicating protein denaturation. Enzyme concentrations in the reaction were ~0.56 µM. Mean ± S.D. of triplicate measurements are shown. The accuracy for the actual temperature in the samples at 50 ˚C was +-2.1 ˚C, but was not measured for the other two temperatures.

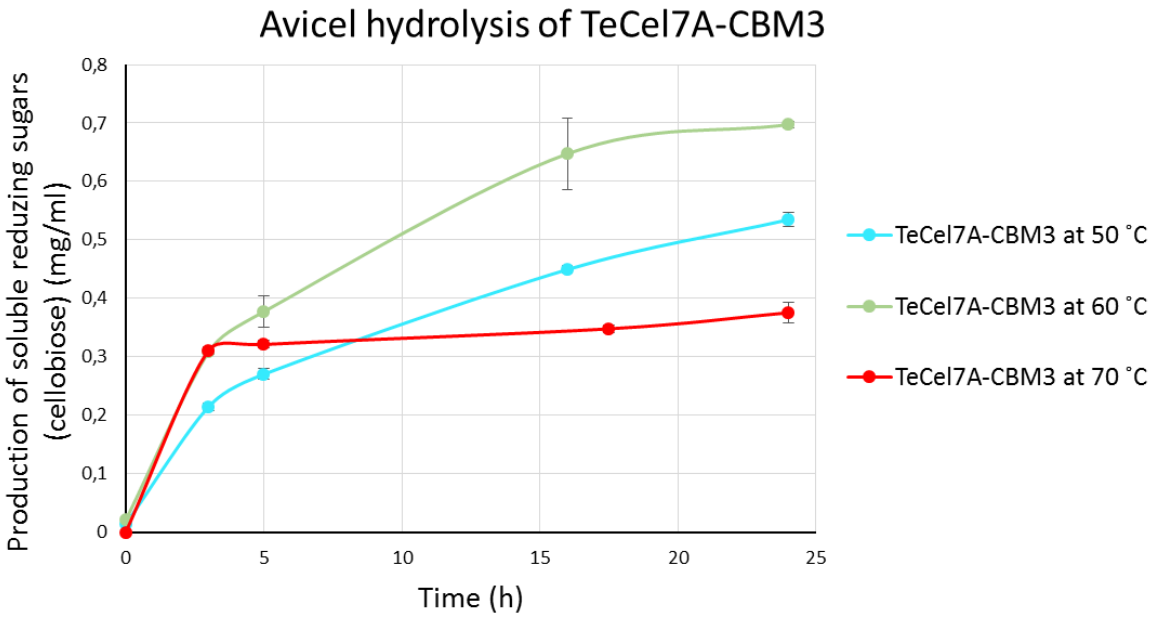


Figure 28. The Avicel (1%) hydrolysis ratios of the *Saccharomyces cerevisiae* produced TeCel7A-CBM3 fusion at different temperatures. The Avicel hydrolysis rate of the fusion enzyme is enhanced by the rise in temperature from 50 ˚C (blue) to 60 ˚C (green) and the best hydrolysis rates are achieved at 60 ˚C. The hydrolysis rate at 70 ˚C (red) is approximately half of the rate at 60 ˚C. The hydrolysis does not proceed at 70 ˚C after 3 hours, at which time the rates are at the same levels as with the 60 ˚C hydrolysis. Enzyme concentrations in the reaction were ~0.56 µM. Mean ± S.D. of triplicate measurements are shown. The accuracy for the actual temperature in the samples at 50 ˚C was +-2.1 ˚C, but was not measured for the other two temperatures.

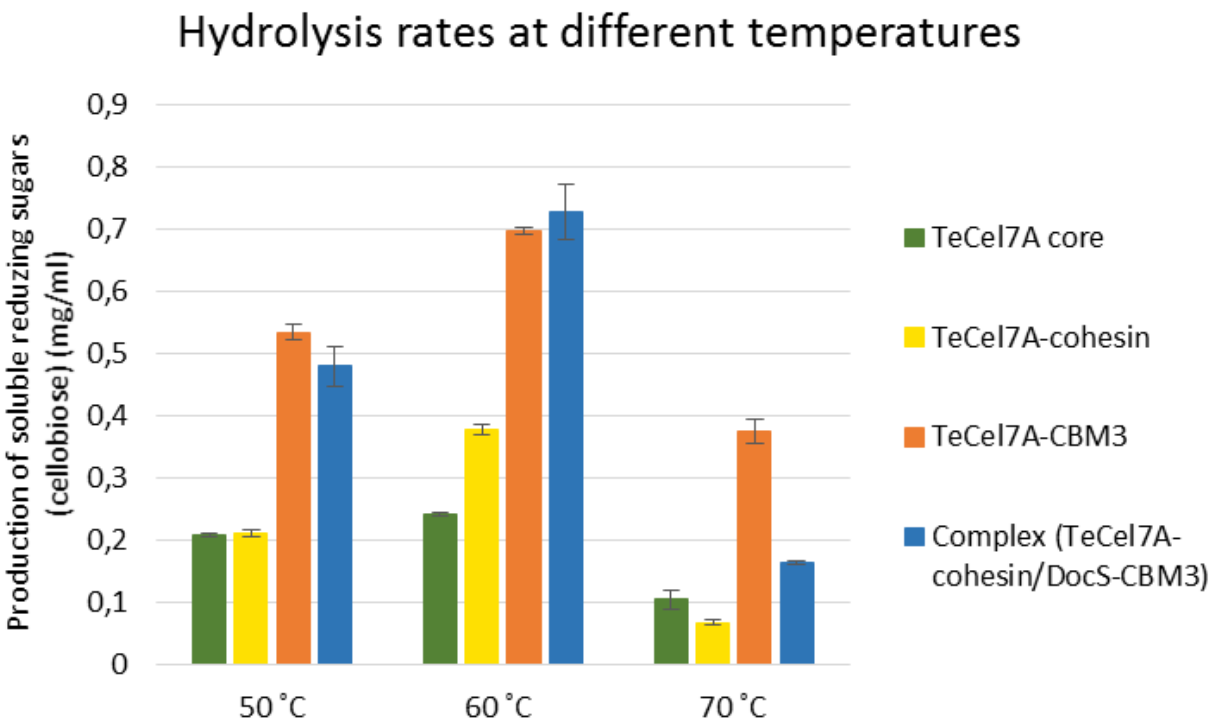


Figure 29. Avicel (1%) hydrolysis rates at different temperatures. Hydrolysis rates depicted here are the amounts of soluble reducing sugars produced by the different enzymes after 24 hours of incubation, although at 70 ˚C the hydrolysis did not proceed further after 3 hours of incubation. The highest hydrolysis rates for all of the enzymes are acquired at 60 ˚C. At 70 ˚C the hydrolysis rates have dropped with varying ratios with different enzymes compared to the ratios at 60 ˚C, with the complex (blue) and the TeCel7A-cohesin fusion (yellow) being the most affected. Enzyme concentrations in the reaction were ~0.56 µM. Mean ± S.D. of triplicate measurements are shown. The accuracy for the actual temperature in the samples at 50 ˚C was +-2.1 ˚C, but was not measured for the other two temperatures.

4. Discussion

4.1. Hypothesis evaluation

In this thesis work the heterologous expression of several different dockerin and CBM containing fusion proteins (Table 2) was attempted in both *E. coli* and *S. cerevisiae* (section 3.1) and the functionality of these proteins in a complex with the TeCel7A catalytic domain-cohesin fusion protein in crystalline cellulose hydrolysis was tested (section 3.2). The best results were obtained with the DocS-CBM3 fusion construct 1 (see Table 2) produced in *E. coli*, having a His-tag in the N-terminus (the dockerin end) of the protein. Characterizations of the cohesin-dockerin linked TeCel7A-CBM3 enzyme complex (including thermostability measurements and activity on soluble and insoluble substrate) showed that the complex performs similarly to the corresponding directly linked fusion enzyme (control enzyme) at temperatures of 50 ˚C and 60 ˚C. At 70 ˚C the complex did not perform as well as the control enzyme, apparently due to instability of the cohesin-dockerin

interaction. The thermostability measurements of the enzymes, together with the previously published data, supported these results.

The functionality of the method and the original hypothesis was, thus, in part verified in this Masters' thesis. The thermostability of the system would need to be improved if screening for enzymes at high temperatures is the goal.

The biggest hurdles encountered in the work were in the heterologous expression of a functional dockerin-CBM1 containing proteins in sufficient amounts. The different *E. coli* strains used in this work could not seemingly produce a functional dockerin-CBM1 fusion protein. The production of a functional CBM1 containing dockerin-CBM fusion protein was, however, achieved in *S. cerevisiae*, although only in poor production yields.

On the other hand, production of functional dockerin-CBM3 fusion proteins in sufficient amounts was successful in *E. coli*. Five constructs, containing five different linker peptides were tested for the DocS-CBM3 fusions, with no obvious effects either on the production levels or the proteolysis. Construct 1 with its N-terminal (dockerin end) His-tag enabled a fast and easy purification of the full length protein, while with the other constructs (with C-terminal His-tags) several smaller proteins were also present in the purification elution fractions. In general, the expression, purification and subsequent handling of the His-tagged DocS-CBM3 fusion proteins was quite fast and straightforward. The proteins could be stored at -20 ˚C and frozen and thawed several times without significant changes in their function. Complex formation was shown to happen within 10 minutes in a buffer solution supplemented with 5 mM $CaCl_2$. The His-tag located at the N-terminal end of the dockerin domain did not seem to disrupt the binding of the cohesin and dockerin domains. Using His-tags for the purification seems like a good strategy for the further studies with these proteins (and possibly also with the CBH-cohesin fusions). Furthermore, it would seem advisable to apply the His-tags distally at the dockerin end of the proteins to enable the purification of only the full length proteins.

4.2. Troubleshooting

The main challenges considering the execution of the method in this thesis work were in the production yields and proteolysis of the dockerin-CBM fusion proteins and in the cellulose hydrolysis with the protein complex at temperatures of 70 ˚C (or above 60 ˚C).

All of the dockerins (even the longer ones) in the DocD-CBM1 –constructs used in this thesis work were shortened or truncated versions of the native Ct endoglucanase D dockerins and none of them contained the residues needed for the intramolecular clasp formation which has been found to confer stability to the dockerin domains (Figure 6)(Slutzki et al. 2013, Chen et al. 2014). Since the clasp forming residues are found in all of the dockerin parts of the DocS-CBM3 fusion constructs (tyrosine 7/13 and proline 68/74, Supplementary files 6 and 7), which were all found at much greater levels in the soluble fraction of the cell lysates, the absence of them in the DocD-CBM1 fusions might be a possible reason why these proteins were found in such low levels in the *E. coli* cell lysates and *S. cerevisiae* culture supernatant.

The results of the production levels between the different versions of the DocD-CBM1 fusions in *E. coli* and *S. cerevisiae* systems showed that the production levels for the two different protein constructs in the two organisms were opposite to each other. This could imply that the leader sequence (*Trichoderma reesei* xylanase 2 leader) in the shorter version of the *S. cerevisiae* produced proteins (which was produced weaker in *E. coli* but was the better one in *S. cerevisiae*) would be a much better choice as a leader sequence than the yeast alpha factors to enable higher yields for the proteins in *S. cerevisiae*.

Another interesting finding was the differential glycosylation patterns of the yeast produced fusion proteins in buffered and unbuffered media. The two putative N-glycosylation sites of the dockerin (NST and NSS sequences existing in both of the duplicated calcium binding loops) had been mutated (to NSA) in the *S. cerevisiae* expressed DocD-CBM1 fusions, but the proteins still possessed putative O-glycosylation sites for the *S. cerevisiae*, especially in the linker region. Buffering of the yeast cultivation media resulted in what seemed like a less glycosylated form of the DocD-CBM1 fusion protein, and buffering of the media should probably be applied also in the future when producing these proteins in yeast.

The thermostability of the cohesin-dockerin interaction was supposedly the reason for the low cellulose hydrolysis activity of the enzyme complex at 70 ˚C. Enzyme complexes containing the *C. thermocellum* Cel48S dockerin (CelS) with its cohesin counterpart have previously been reported to withstand more than 48 hours at 60 ˚C temperature but only some 3 hours at 70 ˚C (Moraïs et al. 2016). Thus it is possible that the inherent thermostability or affinity of the interaction between this particular native cohesin and dockerin pair may not be suitable for hydrolysis at such high temperatures for long periods of time. Improving the affinity and thermostability of the cohesin-

dockerin interaction should therefore be considered a high priority when designing further protein constructs. There are many possibilities to enhance the thermostability of the different domains and possibility to use cohesins and dockerins from hyperthermophilic organisms could also be considered (Moraïs et al. 2016).

4.3. Significance of the results

The results obtained in this thesis are in line with the findings obtained by Vazana et al. (2010) concerning the hydrolysis capability of cohesin-dockerin linked enzymes as compared to directly linked enzymes. Vazana et al. reported that the directly linked enzymes (processive *C. thermocellum* endoglucanases and exoglucanases (Cel9I/Cel9R and Cel48Y/Cel48R)) in which CBMs were attached to the catalytic domains via peptide linkers behaved essentially similarly in degrading cellulose as the corresponding enzymes with the cohesin-dockerin complex linkage, and that the catalytic activity of the enzymes was not ”significantly impaired” by turning the native CBM bearing enzymes into the cellulosomal mode. In this thesis work, the results suggested that there would not be any difference between the intact or cohesin-dockerin linked multidomain enzymes at temperatures of 60˚C or lower. In the study by Vazana et al. (2010) the cellulose hydrolysis capability between the different enzyme systems was only compared at one temperature so no comparisons of the thermal behavior of the both systems could be derived from the experiment, unlike in this thesis work.

It has been proposed that not all cellulase catalytic domains may be suitable to work in a “cellulosomal mode” as are the proteins analogous to those naturally found in cellulosomes, including enzymes from GH families 48, 9, 5, 10, 11 and 26(Caspi et al. 2008, Artzi et al. 2017). In this thesis the GH family 7 cellobiohydrolase of *Talaromyces emersonii* (TeCel7A) was successfully turned into cellulosomal mode by fusing it with a cohesin domain and bringing it into a complex with a DocS-CBM3 fusion protein to enhance its hydrolytic activity to the same level with its corresponding directly linked version at temperatures of 60 ˚C and below.

4.4. Further studies proposals

Further studies should focus on tackling both the proteolysis problems and the low expression levels of some the proteins (DocD-CBM1 fusion constructs). Simultaneously attention should also be paid to the thermostability of the complex and especially the dockerin part of the proteins, which also very well may go hand in hand with the other features. The production of full length dockerins with the clasp forming residues would seem to be a good starting point to try to enhance the production

yields and to prevent excess proteolysis. Using different protease inhibition cocktails such as e.g. a mixture of 1 mM phenylmethylsulfonyl fluoride, 0.4 mM benzamidine and 0.06 mM benzamide (Caspi et al. 2006) in the cell lysis and elution solutions might be at least initially attempted to reduce the amount of proteolysis. Since the fusion proteins produced here were thermostable at least up to 60 ˚C, even a heat denaturing step at these high temperatures could be used to try to denature the *E. coli* proteases before or after cell lysis (Vazana et al. 2010). As the thermostability studies of the proteins with CD showed differential folding of the DocS-CBM3 fusion in the absence and presence of Ca^2, calcium should always be present when either expressing or working with dockerin containing proteins. Furthermore it might also be worthwile to try to produce these proteins in *E. coli* as periplasmic expression. Some other protease deficient or eukaryotic disulfide bridge forming strains of *E. coli* could also be used to try overcome the proteolysis and the CBM1 folding problems.

Since the His-tags did not seem to disrupt the interaction between the dockerin and the cohesin they should be used in the purification of the proteins to enable fast purification of only the full length proteins. It should probably at some point be tested whether or not the His-tag actually interferes with the binding and the affinity of the cohesin and dockerin to each other, especially if the thermostability of the interaction is not increased by other means to sufficiently high levels.

To increase the overall thermostability of the interaction, the use of dockerins and their corresponding cohesins from hyperthermophilic organisms such as *Archaeoglobus fulgidus* or *Thermotoga maritima* could be attempted (Moraïs et al. 2016). Furthermore, mutations in the clasp forming residues of the dockerin e.g. to cysteine residues to try to form disulfide bridges might be attempted to increase the domains thermostability.

Since the production of functional CBM1s was not achieved in *E. coli*, they need to be produced in *S. cerevisiae* or other production hosts. To further increase the production levels in yeast, the expression constructs and subsequently the proteins, in addition to the aforementioned features, should be chosen and designed in a way so that they do not contain yeast N-glycosylation sites (especially in the ligand recognizing sites), the amount of O-glycosylation sites in the linker part is minimal and they contain highly efficient signal sequences (e.g. initially the *T. reesei* xylanase 2 signal sequence) to maximize the yields.

Finally, more investigations could also be carried out with different types of enzymes on different configurations and with different substrates to confirm the broader applicability of the method. The method could also be tested as a general method for cellulase protein engineering.

Acknowledgement:

I want to thank my supervisors Sanni Voutilainen, Anu Koivula and Kiyohiko Igarashi from VTT for their time and expertise in guiding me through this project. Professor Marko Virta from the University of Helsinki I want to thank for his comments in the finalization stage of the thesis.

I also want to thank everybody else at VTT with whom I worked with for their kind help and assistance with all sorts of problems and for keeping the spirit up high during the thesis work.

Last, I would very much like to thank my family and friends for their support throughout the thesis work and my studies on and off the subject and especially I want to thank Annaleena and Venla for constantly making my days cheerful and full of laughter and play.

References:

Anton, B., Formenkov, A., Raleigh, E. & Berkmen, M. 2016. Complete genome sequence of the engineered SHuffle strains and their wild-type parents. Genome Announcements 4: 2. American Society for Microbiology.

Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. 2006. The SWISS-MODEL Workspace: A web based environment for protein structure homology modelling. Bioinformatics 22: 195-201.

Artzi, L., Bayer, E. & Moraïs, S. 2017. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. Nature Reviews Microbiology 15: 83-95.

Barak Y., Handelsman T., Nakar, D., Mechaly, A., Lamed, R., Shoham, Y. & Bayer, E. 2005. Matching fusion protein systems for affinity analysis of two interacting families of proteins: the cohesin-dockerin interaction. Journal of Molecular Recognition 18: 491-501.

Bayer, E. 2017. Cellulosomes and designer cellulosomes: why toy with Nature? Environmental Microbiology Reports 9: 14-15.

Bayer, E., Morag, E. & Lamed, R. 1994. The cellulosome – a treasure-trove for biotechnology. Trends in Biotechnology 12: 379-386.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalow, I. & Bourne, B. 2000. The Protein Data Bank. Nucleic Acids Research 28: 235–242.

Boraston, A., Bolam, D., Gilbert, H. & Davies, G. 2004. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochemical Journal 382: 769-781.

Carrard, G., Koivula, A., Söderlund, H. & Béguin, P. 2000. Cellulose-binding domains promote hydrolysis of different sites on crystalline cellulose. PNAS 97: 10342-10347.

Carvalho, A., Dias, F., Prates, J., Nagy, T., Gilbert, H., Davies, G., Ferreira, L., Romão, M., & Fontes, C. 2003. Cellulosome assembly revealed by crystal structure of the cohesin-dockerin complex. PNAS 100: 13809-13814.

Carvalho, A., Dias, F., Nagy, T., Prates, J., Proctor, M., Smith, N., Bayer, E., Davies, G., Ferreira, L., Romão, M., Fontes, C. & Gilbert, H. 2007. Evidence for a dual binding mode of dockerin modules to cohesins. PNAS 104: 3089–3094.

Caspi, J., Irwin, D., Lamed, R., Li, Y., Fierobe, H-P., Wilson, D. & Bayer, E. 2008. Conversion of *Thermobifida fusca* free exoglucanases into cellulosomal components: Comparative impact on cellulose-degrading activity. Journal of Biotechnology 135: 351–357.

Caspi, J., Irwin, D., Lamed, R., Shoham, Y., Fierobe, H-P., Wilson, D. & Bayer, E. 2006. *Thermobifica fusca* family-6 cellulases as potential designer cellulosome component. Biocatalysis and Biotransformation 24: 3-12.

Chen, C., Cui, Z., Xiao, Y., Cui, Q., Smith, S., Lamed, R., Bayer, E. & Feng, Y. 2014. Revisiting the NMR solution structure of the Cel48S type-I dockerin module from *Clostridium thermocellum* reveals a cohesin-primed conformation. Journal of Structural Biology 188: 188-193.

Craig, S., Foong, F. & Nordon, R. 2006. Engineered proteins containing the cohesin and dockerin domains from *Clostridium thermocellum* provides a reversible, high affinity interaction for biotechnology applications. Journal of Biotechnology 121: 165-173.

Currie, M., Adams, J., Faucher, F., Bayer, E., Jia, Z. & Smith, S. 2012. Scaffoldin Conformation and Dynamics Revealed by a Ternary Complex from the *Clostridium thermocellum* Cellulosome. The Journal of Biological Chemistry 287: 26953-26961.

Divne, C., Ståhlberg, J., Reinikainen, T., Ruohonen, L., Petterson, G., Knowles, J., Teeri, T. & Jones, T. 1994. The Three-Dimensional Crystal Structure of the Catalytic Core of Cellobiohydrolase I from *Trichoderma reesei*. Science 265: 524-528.

Fanutti, C., Ponyi, T., Black, G., Hazlewood, G. & Gilbert, H. 1995. The Conserved Noncatalytic 40-Residue Sequence in Cellulases and Hemicellulases from Anaerobic Fungi Functions as a Protein Docking Domain. The Journal of Biological Chemistry 270: 29314-29322.

Fierobe, H-P., Gaudin, C., Belaich, A., Loutfi, M., Faure, M., Bagnara, C., Baty, D. & Belaich, J-P. 1991. Characterization of Endoglucanase A from *Clostridium cellulolyticum*. Journal of Bacteriology 173: 7956-7962.

Florin, E., Moy, V. & Gaub, H. 1994. Adhesion forces between individual ligand receptor pairs. Science 264: 415-417**.**

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M., Appel, R. & Bairoch, A. 2005. Protein Identification and Analysis Tools on the ExPASy Server. In: John M. Walker (ed.), The Proteomics Protocols Handbook, Humana Press, pp. 571-607.

Gietz, R. & Woods, R. 2002 Transformation of yeast by lithium acetate/ single-stranded carrier DNA/ polyethylene glycol method. Methods in Molecular Biology 350: 87-96.

Gilbert, H. & Hazlewood, G. 1993. Bacterial cellulases and xylanases. Journal of General Microbiology 139: 187-194.

Gerngross, U., Romaniec, M., Huskisson, N. & Demain, A. 1993. Sequencing of *Clostridium thermocellum* gene (cipA) encoding the cellulosomal S_L-protein reveals an unusual degree of internal homology. Molecular Microbiology 8: 325-334.

Grassick, A., Murray, P., Thompson, R., Collins, C., Byrnes, L., Birrane, G., Higgins, T. & Tuohy, M. 2004. Three-dimensional structure of a thermostable native cellobiohydrolase, CBH IB, and molecular characterization of the cel7 gene from the filamentous fungus, *Talaromyces emersonii*. European Journal of Biochemistry 271: 4495–4506.

Guex, N. & Peitsch, M. 1997. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 18: 2714-2723.

Guillén, D., Sánchez, S. & Rodríguez-Sanoja, R. 2010. Carbohydrate-binding domains: multiplicity of biological roles. Applied Microbiology and Biotechnology 85: 1241-1249.

Hahm, J., Znameroski, E., Liu, F., Heu, T., Haydon, I., Hasani, S., Lamsa, M., Jones, A., Widner, W., Mullikin, R., Harris, P., Teter, S. & Lin, J. 2015. Development and application of a synthetic cellulosome-based screening platform for enhanced enzyme discovery. Poster abstract, 11th Carbohydrate Bioengineering Meeting, 10.-13.5.2015, Espoo, Finland.

Jobst, M., Milles, L., Schoeler, C., Ott, W., Fried, D., Bayer, E., Gaub, H. & Nash, M. 2015. Resolving dual binding conformations of cellulosome cohesin-dockerin complexes using single-molecule force spectroscopy. eLife 2015;4:e10319.

Karpol, A., Barak, Y., Lamed, R., Shoham, Y. & Bayer, E. 2008. Functional asymmetry in cohesin binding belies inherent symmetry of the dockerin module: insight into cellulosome assembly revealed by systematic mutagenesis. Biochemical Journal 410: 331-338.

Kim, S. & Hahn, J-S. 2014. Synthetic scaffold based on a cohesin–dockerin interaction for improved production of 2,3-butanediol in *Saccharomyces cerevisiae*. Journal of Biotechnology 192: 192-196.

Kelley, L., Mezulis, S., Yates, C., Wass, M. & Sternberg, M. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols 10: 845–858.

Knowles, J., Lentovaara, P., Murray, M. & Sinnot, M. 1988. Stereochemical Course of the Action of the Cellobioside Hydrolases I and II of *Trichoderma reesei*. Journal of the Chemical Society, Chemical Communications 1988: 1401-1402.

Kraulis, J., Clore, G., Nilges, M., Jones, T., Pettersson, G., Knowles, J. & Gronenborn, A. 1989. Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. Biochemistry 28: 7241-7257.

Lamed, R., Setter, E. & Bayer, E. 1983. Characterization of a Cellulose-Binding, Cellulase-Containing Complex in *Clostridium thermocellum*. Journal of Bacteriology 156: 828-836.

Lever, M. 1972. A New Reaction for Colorimetric Determination of Carbohydrates. Analytical Biochemistry 47: 273-279.

Lytle, B., Myers, C., Kruus, K. & Wu, J. 1996. Interactions of the CelS Binding Ligand with Various Receptor Domains of the *Clostridium thermocellum* Cellulosomal Scaffolding Protein, CipA. Journal of Bacteriology 178: 1200-1203.

Lytle, B., Volkman, B., Westler, W. & Wu, J. 2000. Secondary Structure and Calcium-Induced Folding of the *Clostridium thermocellum* Dockerin Domain Determined by NMR Spectroscopy. Archives of Biochemistry and Biophysics 379: 237-244.

Mechaly, A., Fierobe, H-P., Belaich, A., Belaich, J-P., Lamed, R., Shoham, Y. & Bayer, E. 2001. Cohesin-Dockerin Interaction in Cellulosome Assembly. A Single Hydroxyl Group of a Dockerin Domain Distinguishes Between Nonrecognition and High Affinity Recognition. The Journal of Biological Chemistry 276: 9883-9888.

Mechaly, A., Yaron, S., Lamed, R., Fierobe, H-P., Belaich, A., Belaich, J-P., Shoham, Y. & Bayer, E. 2000. Proteins: Structure, Function, and Genetics 39: 170-177.

Merkel, R., Nassoy, P., Leung, A., Ritchie, K. & Evans, E. 1999. Energy landscapes of receptor-ligand bonds explored with dynamic force spectroscopy. Nature 397: 50-53.

Miras, I., Schaeffer, F., Béguin, P. & Alzari, P. 2002. Mapping by Site-Directed Mutagenesis of the Region Responsible for Cohesin-Dockerin interaction on the Surface of the Seventh Cohesin Domain of *Clostridium thermocellum* CipA. Biochemistry 41: 2115-2119.

Moraïs, S., Stern, J., Kahn, A., Galanopoulou, A., Yoav, S., Shamshoum, M., Smith, M., Hatzinikolaou, D., Arnold, F. & Bayer, E. 2016. Enhancement of cellulosome-mediated deconstruction of cellulose by improving enzyme thermostability. Biotechnology for Biofuels 9: 164.

Murashima, K., Chen, C-L., Kosugi, A., Tamaru, Y., Doi, R. & Wong S-L. 2001. Heterologous production of *Clostridium cellulovorans* engB, Using Protease-Deficient *Bacillus subtilis*, and Preparation of Active Recombinant Cellulosomes. Journal of Bacteriology 184: 76-81.

National Renewable Energy Laboratory (NREL). 2018. Gregg Beckham. https://www.nrel.gov/research/gregg-beckham.html. Read and printed 16.2.2018.

Pagès, S., Bélaïch, A., Bélaïch, J-P., Morag, E., Lamed, R., Shoham, Y. & Bayer, E. 1997. Species-Specifity of the Cohesin-Dockerin Interaction Between *Clostridium thermocellum* and *Clostridium cellulolyticum*: Prediction of Specificity Determinants of the Dockerin Domain. Proteins 29: 517-527.

Payne, C., Knott, B., Mayes, H., Hansson, H., Himmel, M., Sandgren, M., Ståhlberg, J. & Beckham, G. 2015. Fungal Cellulases. Chemical Reviews 115: 1308-1448.

Peränen, J., Rikkonen, M., Hyvönen, M. & Kääriäinen, L. 1996. T7 Vectors with a Modified T7lac Promoter for Expression of Proteins in *Escherichia coli*. Analytical Biochemistry 236: 371–373.

Schoeler, C., Malinowska, K., Bernardi, R., Milles, L., Jobst, M., Durner, E., Ott, W., Fried, D., Bayer, E., Schulten, K., Gaub, H. & Nash, M. 2014. Ultrastable cellulosome-adhesion complex tightens under load. Nature Communications 5: 5635.

Shimon, L., Bayer, E., Morag, E., Lamed, R., Yaron, S., Shoham, Y. & Frolow, F. 1997. A cohesin domain from *Clostridium thermocellum*: the crystal structure provides new insights into cellulosome assembly. Structure 5: 381- 390.

Slutzki, M., Jobby, M., Chitayat, S., Karpol, A., Dassa, B., Barak, Y., Lamed, R., Smith, S. & Bayer, E. 2013. Intramolecular clasp of the cellulosomal *Ruminococcus flavefaciens* ScaA dockerin module confers structural stability. FEBS Open Bio 3: 398-405.

Smith, S. & Bayer, E. 2013. Insights into cellulosome assembly and dynamics: from dissection to reconstruction of the supramolecular enzyme complex. Current Opinion in Structural Biology 23: 686-694.

Stahl, S., Nash, M., Fried, D., Slutzki, M., Barak, Y., Bayer, E. & Gaub, H. 2012. Single-molecule dissection of the high-affinity cohesin-dockerin complex. PNAS 190: 20431-20436.

Ståhlberg, J., Divne, C., Koivula, A., Piens, K., Claeyssens, M., Teeri, T. & Jones, A. 1996. Activity studies and Crystal Structure of Catalytically Deficient Mutants of Cellobiohydrolase I from *Trichoderma reesei*. Journal of Molecular Biology 264: 337-349.

The Architecture et Fonction des Macromolécules Biologiques (AFMB) laboratory. 2017. The Carbohydrate-Active Enzymes database (CAZy). http://www.cazy.org. Updated 10.7.2017. Read and printed 12.7.2017.

Várnai, A., Mäkelä, M., Djajadi, D., Rahikainen, J., Hatakka, A. & Viikari, L. 2014. Chapter Four - Carbohydrate-Binding Modules of Fungal Cellulases: Occurrence in Nature, Function, and Relevance in Industrial Biomass Conversion. Advances in Applied Microbiology 88: 103-165.

Vazana, Y., Moraïs, S., Barak, Y., Lamed, R. & Bayer, E. 2010. Interplay between *Clostridium thermocellum* Family 48 and Family 9 Cellulases in Cellulosomal versus Noncellulosomal States. Applied and Environmental Microbiology 76: 3236-3243.

Viikari, L., Vehmaanperä, J. & Koivula, A. 2012. Lignocellulosic ethanol: From science to industry. Biomass and Bioenergy 46: 13-24.

Voutilainen, S., Nurmi-Rantala, S., Penttilä, M. & Koivula, A. 2014. Engineering chimeric thermostable GH7 cellobiohydrolases in *Saccharomyces cerevisiae*. Applied Microbiology and Biotechnology 98: 2991-3001.

Wilchek, M. & Bayer, E. 1999. Foreword and introduction to the book (strept)avidin–biotin system. Biomolecular Engineering 16: 1-4.

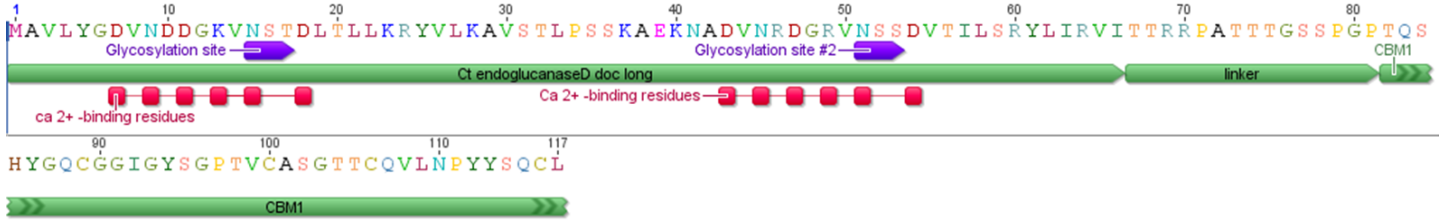
Yaron, S., Morag, E., Bayer, E., Lamed, R. & Shoham, Y. 1995. Expression, purification and subunit-binding properties of cohesins 2 and 3 of the *Clostridium thermocellum* cellulosome. FEBS Letters 360: 121-124.

Zhang, Y-H., Himmel, M. & Mielenz, J. 2006. Outlook for cellulase improvement: Screening and selection strategies. Biotechnology Advances 24: 452-481.

Supplementary files:

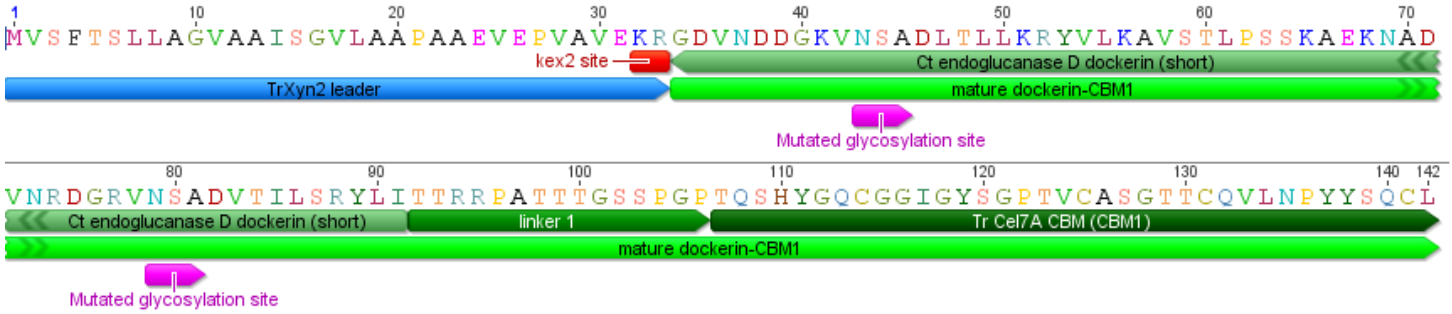
Supplementary file 1. *Clostridium thermocellum* Endoglucanase D dockerin (short)-linker 1-*Trichoderma reesei* cellobiohydrolase I CBM fusion protein (short DocD-CBM1) amino acid sequence from pSVFDP5. Depicted with the green arrows are the different parts and domains of the fusion protein. The sequences for methionine (M) and alanine (A) residues are added to the N-terminal side of the proteins because the sequences coding for these amino acids are needed for translation and in the cloning of the genes to the expression vectors and they remain in the mature protein also. The CBM1 part corresponds to the *T. reesei* CBHI sequence and structure by Kaulis et al. (1989) and the linker is a part of the native *T. reesei* linker sequence preceding that (GenBank ID P62694.1). The sequence for the endoglucanase D dockerin is a truncated version of the Ct endoglucanase D dockerin of PDB ID 1CLC_A and is the shorter of the two versions used in this thesis work.

Supplementary file 2. *Clostridium thermocellum* Endoglucanase D dockerin-linker 1-*Trichoderma reesei* cellobiohydrolase I CBM fusion protein (DocD-CBM1) amino acid sequence from pSVFDP6. Depicted are also the different parts and domains of the fusion protein (green arrows), putative *S. cerevisiae* N-glycosylation sites (NST/NSS, purple arrows) and the Ca^{2+}-binding residues of the dockerin domain (pink squares, Carvalho et al. 2003). The sequences for methionine (M) and alanine (A) residues are added to the N-terminal side of the proteins because the sequences coding for these amino acids are needed for translation and in the cloning of the genes to the expression vectors and they remain in the mature protein also. The CBM1 and linker sequences are the same as describes in supplementary file 1. The sequence for the endoglucanase D dockerin is a truncated version of the Ct endoglucanase D dockerin of PDB ID 1CLC_A and is the longer of the two versions used in this thesis work.

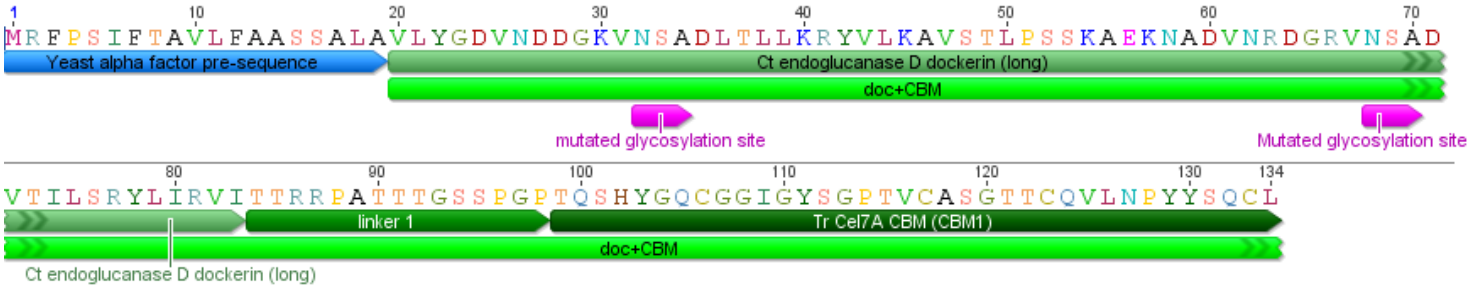


Supplementary file 3. *Talaromyces emersonii* Cel7A catalytic domain-linker 6-*Clostridium thermocellum* CipA cohesin2 fusion protein amino acid sequence from pSVFDP14. Depicted in the figure are the leader sequence (red arrow), and the linker and Ct coh 2 sequences (green upper arrows). The rest of the protein depicted with only one green arrow is the TeCel7A catalytic domain. In the mature protein the leader sequence is cleaved off after the alanine (A) residue, leaving a mature protein with a glutamine (Q) residue in the N-terminus. The sequences of the fusion protein correspond to those of: GenBank ID: AGU68249.1 By Voutilainen et al. (2014): TeCel7A from 1 to 453 with the exception that residue 285 has been mutated to D instead of N to remove a putative N-glycosylation site, linker: a partial linker region of *Trichoderma reesei* cellobiohydrolase 7A (CBHI) as a duplication GenBank ID: P62694.1, cohesin: GenBank ID 1OHZ_A by Carvalho et al. (2003 and 2007) the second cohesin domain of CipA of *C. thermocellum*, with the exception of the truncation of the first 4 residues (M,A,S and D) of 1OHZ_A.

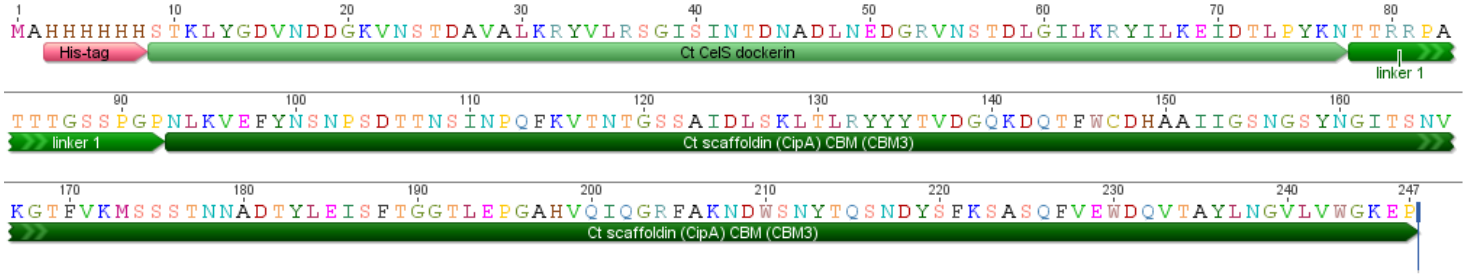
Supplementary file 4. *Clostridium thermocellum* endoglucanase D dockerin (short)-linker 1-*Trichoderma reesei* cellobiohydrolase I CBM (CBM1) fusion protein (short DocD-CBM1) amino acid sequence from pSVFDP12. Depicted in the figure are the *Trichoderma reesei* Xylanase 2 leader sequence (blue arrow), the mature fusion protein (lime green arrow), the dockerin (light green arrow), the linker (green arrow) and the CBM1 (dark green arrow) parts of the protein. In the mature protein the leader sequence is cleaved off after the kex2 protease site (red block), leaving a mature protein with a glysine (G) residue in the N-terminus. The mutated putative *Saccharomyces cerevisiae* N-glycosylation sites are depicted with pink arrows. The dockerin part of the fusion protein is truncated by 6 amino acids, compared to the longer version of the protein (supplementary file 2 and 5). The valine (V), leusine (L) and tyrosine (Y) residues from the N-terminal end of the dockerin part and the arginine (R), valine (V), and isoleucine (I) residues from the C-terminal end of the dockerin part are missing, when compared to the longer protein. The CBM1 and linker sequences are the same as describes in supplementary file 1. The sequence for the endoglucanase D dockerin is a truncated and the shorter of the two versions of the Ct endoglucanase D dockerin of PDB ID 1CLC_A used in this thesis work and essentially the same as in supplementary file 1 besides the putative N-glycosylation site mutations in residues 45 and 81 (purple arrow).



Supplementary file 5. *Clostridium thermocellum* endoglucanase D dockerin (long)-linker 1-*Trichoderma reesei* cellobiohydrolase I CBM (CBM1) fusion protein (DocD-CBM1) amino acid sequence from pSVFDP13. Depicted in the figure are the Yeast alpha factor pre-sequence (blue arrow), the mature fusion protein (lime green bar), the dockerin (light green arrow), the linker (green arrow) and the CBM1 (dark green arrow) parts of the protein. In the mature protein the leader sequence is cleaved off after the alanine (A) residue, leaving a mature protein with a valine (V) residue in the N-terminus. The mutated putative *Saccharomyces cerevisiae* N-glycosylation sites are depicted with pink arrows. The CBM1 and linker sequences are the same as describes in supplementary file 1. The sequence for the endoglucanase D dockerin is a truncated, but longer of the two versions of the Ct endoglucanase D dockerin of PDB ID 1CLC_A used in this thesis work and essentially the same as in supplementary file 2 besides the putative N-glycosylation site mutations in residues 34 and 70 (purple arrow).



Supplementary file 6. *Clostridium thermocellum* CelS dockerin (DocS)-linker 1-*Clostridium thermocellum* scaffoldin (CipA) CBM (CBM3) fusion protein amino acid sequence from pSVFDP21. Depicted in the figure are the N-terminal Histidine-tag (light red), the DocS (light green arrow), the linker (green arrow) and the CBM3 (dark green arrow) parts of the protein. The sequences for methionine (M) and alanine (A) residues are added to the N-terminal side of the protein because the sequences coding for these amino acids are needed for translation and in the cloning of the gene to the expression vector and they remain in the mature protein also. The dockerin corresponds to the dockerin part of the Doc48S (PDB ID 2MTE_A). The linker part is a partial *T. reesei* cellobiohydrolase 7A (CBHI) linker (GenBank ID P62694.1) and is the same as with DocD-CBM1 –constructs. The CBM3 part of the protein correspond to that of PDB ID 1NBC_A.



Supplementary file 7. *Clostridium thermocellum* CelS dockerin (DocS)-linker (2 to 4)-*Clostridium thermocellum* scaffoldin (CipA) CBM (CBM3) fusion proteins' amino acid sequences from pSVFDP22 to pSVFDP25. Depicted in the figures are the C-terminal Histidine-tag (light red), the DocS (light green arrow), the linker (green arrow) and the CBM3 (dark green arrow) parts of the proteins. The sequences for methionine (M) and alanine (A) residues are added to the N-terminal side of the proteins because the sequences coding for these amino acids are needed for translation and in the cloning of the genes to the expression vectors and they remain in the mature protein also. The dockerin part of the protein is the dockerin part of the Doc48S (PDB ID 2MTE_A) and is the same as in the DocS-CBM3 – construct 1. The linkers in these proteins are of bacterial origin unlike the linker 1 which is of a fungal origin. Linkers 2, 3 and 4 are from GenBank accession numbers AF283514, AF283515 and AF283517 respectively (Carrard et al. 2000). Linker 5 is from a *Cellulomonas fimi* Cex (GH10 xylanase)(Sanni Voutilainen, personal communication). The histidine-tags in these proteins are located at the C-termini opposite to the DocS-CBM3 fusion construct 1 in which the histidine tag is located at the N-terminus. The CBM3 part of the protein correspond to that of PDB ID 1NBC_A.